Non-parametric Stochastic Approximation with Large Step sizes

Aymeric Dieuleveut, Francis Bach

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

NON-PARAMETRIC STOCHASTIC APPROXIMATION WITH LARGE STEP SIZES

By Aymeric Dieuleveut, Francis Bach

Département d'Informatique de l'Ecole Normale Supérieure, Paris, France

Abstract We consider the random-design least-squares regression problem within the reproducing kernel Hilbert space (RKHS) framework. Given a stream of independent and identically distributed input/output data, we aim to learn a regression function within an RKHS $\mathcal{H}$, even if the optimal predictor (i.e., the conditional expectation) is not in $\mathcal{H}$. In a stochastic approximation framework where the estimator is updated after each observation, we show that the averaged unregularized least-mean-square algorithm (a form of stochastic gradient), given a sufficient large step-size, attains optimal rates of convergence for a variety of regimes for the smoothnesses of the optimal prediction function and the functions in $\mathcal{H}$.

1. Introduction. Positive-definite-kernel-based methods such as the support vector machine or kernel ridge regression are now widely used in many areas of science of engineering. They were first developed within the statistics community for non-parametric regression using splines, Sobolev spaces, and more generally reproducing kernel Hilbert spaces (see, e.g., [1]). Within the machine learning community, they were extended in several interesting ways (see, e.g., [2, 3]): (a) other problems were tackled using positive-definite kernels beyond regression problems, through the "kernelization" of classical unsupervised learning methods such as principal component analysis or K-means, (b) efficient algorithms based on convex optimization have emerged, and (c) kernels for non-vectorial data have been designed for objects like strings, graphs, measures, etc. A key feature is that they allow the separation of the representation problem (designing good kernels for non-vectorial data) and the algorithmic/theoretical problems (given a kernel, how to design, run efficiently and analyse estimation algorithms).

The theoretical analysis of non-parametric least-squares regression within the RKHS framework is well understood. In particular, for regression on input data in $\mathbb{R}^d$, $d \geqslant 1$, and so-called Mercer kernels (continuous kernels over a compact set) that lead to dense subspaces of the space of square-integrable functions, the optimal rates of estimation given the smoothness

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35
Keywords and phrases: Reproducing kernel Hilbert space, Stochastic approximation

of the optimal prediction function are attained for a sufficiently small Hilbert space of functions [4, 5, 6]. However, the kernel framework goes beyond Mercer kernels and non-parametric regression; indeed, kernels on non-vectorial data provide examples where the usual topological assumptions may not be natural, such as sequences, graphs and measures. Moreover, even finite-dimensional Hilbert spaces may need a more refined analysis when the dimension of the Hilbert space is much larger than the number of observations: for example, in modern text and web applications, linear predictions are performed with a large number of covariates which are equal to zero with high probability. The sparsity of the representation allows to reduce significantly the complexity of traditional optimization procedures; however, the finite-dimensional analysis which ignores the spectral structure of the data is not applicable, while the analysis we carry out is. In this paper, we consider minimal assumptions regarding the input space and the distributions, so that our non-asymptotic results may be applied to all the cases mentioned above.

In practice, estimation algorithms based on regularized empirical risk minimization face two challenges: (a) using the correct regularization parameter and (b) finding an approximate solution of the convex optimization problems. In this paper, we consider these two problems jointly by following a stochastic approximation framework formulated directly in the RKHS, in which each observation is used only once and overfitting is avoided by making only a single pass through the data (a form of early stopping). While this framework has been considered before [7, 8, 9], the algorithms that are considered either (a) require two sequences of hyperparameters (the step-size in stochastic gradient descent and a regularization parameter) or (b) do not always attain the optimal rates of convergence for estimating the regression function. In this paper, we aim to remove simultaneously these two limitations.

Traditional online stochastic approximation algorithms, as introduced by Robbins and Monro [10], lead in finite-dimensional learning problems to stochastic gradient descent methods with steps decreasing with the number of observations n, which are typically proportional to $n^{-\zeta}$, with ζ between $1/2$ and 1. Short step-sizes ($\zeta = 1$) are adapted to well-conditioned problems (low dimension, low correlations between covariates), while longer step-sizes ($\zeta = 1/2$) are adapted to ill-conditioned problems (high dimension, high correlations) but with a worse convergence rate—see, e.g., [11, 12] and references therein). More recently [13] showed that constant steps with averaging could lead to the best possible convergence rate in Euclidean spaces (i.e., in finite dimensions). In this paper, we show that using longer step-sizes with

averaging also brings benefits to Hilbert space settings.

With our analysis, based on positive definite kernels, under assumptions on both the objective function and the covariance operator of the RKHS, we derive improved rates of convergence [5], in both the finite horizon setting where the number of observations is known in advance and our bounds hold for the last iterate (with exact constants), and the online setting where our bounds hold for each iterate (asymptotic results only). It leads to an explicit choice of the learning rates which may be used in stochastic gradient descent, depending on the number of training examples we want to use and on the assumptions we make.

In this paper, we make the following contributions:

- We outline in Section 2 a general though simple algebraic framework for least-squares regression in RKHS, which encompasses all commonly encountered situations, and avoids many of the assumptions that are often made. This allows to cover many practical examples.
- We characterize in Section 3 the convergence rate of averaged least-mean-squares (LMS) and show how the proper set-up of the step-size leads to optimal convergence rates of prediction (as they were proved in [5]), extending results from finite-dimensional [13] to infinite-dimensional settings. The problem we solve here was stated as an open problem in [8, 7].
- We compare our new results with existing work, both in terms of rates of convergence in Section 4, and with simulations on synthetic spline smoothing in Section 5.

2. Learning with positive-definite kernels. In this paper, we consider a general random design regression problem, where observations (x_i, y_i) are independent and identically distributed (i.i.d.) random variables in $\mathcal{X} \times \mathcal{Y}$ drawn from a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ may be any set equipped with a measure, while for simplicity, we only consider $\mathcal{Y} = \mathbb{R}$ and we measure the risk of a function $g : \mathcal{X} \to \mathbb{R}$, by the mean square error, that is, $\varepsilon(g) := \frac{1}{2}\mathbb{E}_\rho\left[(g(X) - Y)^2\right]$.

The function g that minimizes $\varepsilon(g)$ is known to be conditional expectation, that is, $g_\rho(X) = \mathbb{E}[Y|X]$. In this paper we consider formulations where our estimates lie in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this section, we build a general framework which makes "minimal" assumptions regarding the set $\mathcal{X}$ (only assumed to be equipped with a measure), the kernel K (only assumed to have bounded expectation $\mathbb{E}_\rho K(X, X)$) and the output Y (only assumed to have finite variance).

These assumptions include classical examples such as Mercer kernels, finite-dimensional feature spaces, but also apply to kernels on discrete objects (with non-finite cardinality). See examples in Section 2.5.

2.1. *Reproducing kernel Hilbert spaces.* Throughout this paper, we make the following assumption:

(A1) $\mathcal{H}$ is a separable[1] RKHS associated with kernel K on the set $\mathcal{X}$.

RKHSs are well-studied Hilbert spaces which are particularly adapted to regression problems (see, e.g., [14]). They satisfy the following properties:

1. $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a separable Hilbert space of functions: $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$.
2. $\mathcal{H}$ contains all functions $K_x : t \mapsto K(x,t)$, for all x in $\mathcal{X}$.
3. For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the reproducing property holds:

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}.$$

The Hilbert space $\mathcal{H}$ is totally characterized by the positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which simply needs to be a symmetric function on $\mathcal{X} \times \mathcal{X}$ such that for any finite family of points $(x_i)_{i \in I}$ in $\mathcal{X}$, the $|I| \times |I|$-matrix of kernel evaluations is positive semi-definite. We provide examples in Section 2.5.

Note that we do not make any topological assumptions regarding the set $\mathcal{X}$. We will only assume that it is equipped with a probability measure.

2.2. *Random variables.* In this paper, we consider a set $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ and a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. We denote ρ_X the marginal law on the space $\mathcal{X}$ and $\rho_{Y|X=x}$ the conditional probability measure on Y given $x \in \mathcal{X}$. We shall use the notations $\mathbb{E}\left[f(X)\right]$ or $\mathbb{E}_{\rho_X}\left[f(\cdot)\right]$ for $\int_{\mathcal{X}} f(x) d\rho_X(x)$. Beyond the moment conditions stated below, we will always make the assumptions that the space of square-integrable function $\mathcal{L}^2_{\rho_X}$ defined below is separable[2].

Throughout the paper, we make the following simple assumption regarding finitess of moments:

(A2) $\mathbb{E}\left[K(X,X)\right]$ and $\mathbb{E}[Y^2]$ are finite.

In previous work, **(A1)** was typically replaced by the assumption that K is a Mercer kernel ($\mathcal{X}$ compact set and ρ_X with full support, K continuous)[9,

[1]The separability assumption is necessary to be able to expand any element as an infinite sum, using a countable orthonormal family. This assumption is satisfied in almost all cases, for instance it is simple as soon as $\mathcal{X}$ admits a topology for which it is separable and functions in $\mathcal{H}$ are continuous, see [14] for more details.

[2]Like for RKHSs, this is the case in most interesting situations. See [15] for more details.

16, 4, 8], **(A2)** was replaced by the stronger assumptions $\sup_{x\in\mathcal{X}} K(x,x) < \infty$ [9, 8, 7] and $|Y|$ bounded [9, 7]. Note that in functional analysis, the weaker hypothesis $\int_{\mathcal{X}\times\mathcal{X}} k(x,x')^2 d\rho_X(x) d\rho_X(x') < \infty$ is often used [17], but it is not adapted to the machine learning setting.

Our assumptions are sufficient to analyse the minimization of $\varepsilon(f) = \frac{1}{2}\mathbb{E}\left[(f(X)-Y)^2\right]$ with respect to $f \in \mathcal{H}$. In many applications, the minimum is not attained. We now present the usual functional analysis tools necessary to study this minimization problem and we show that the assumptions above are sufficient to carry it through. In this section, we assume given a distribution ρ on the data and we consider exact expectations; we will consider estimation using empirical averages in Section 3.

We first need to introduce the space of square ρ_X-integrable functions $\mathcal{L}^2_{\rho_X}$, and its quotient $L^2_{\rho_X}$ that makes it a separable Hilbert space (see, e.g., [18]). That is,

$$\mathcal{L}^2_{\rho_X} = \left\{ f \ : \mathcal{X} \to \mathbb{R} \ \Big/ \int_{\mathcal{X}} f^2(t) d\rho_X(t) < \infty \right\};$$

moreover $L^2_{\rho_X}$ is the quotient of $\mathcal{L}^2_{\rho_X}$ under the equivalence relation given by

$$f \equiv g \Leftrightarrow \int_{\mathcal{X}} (f(t)-g(t))^2 d\rho_X(t) = 0.$$

We denote by $\|\cdot\|_{\mathcal{L}^2_{\rho_X}}$ the (semi-)norm:

$$\|f\|^2_{\mathcal{L}^2_{\rho_X}} = \int_{\mathcal{X}} |f(x)|^2 d\rho_X(x).$$

The space $L^2_{\rho_X}$ is then a Hilbert space with norm $\|\cdot\|_{L^2_{\rho_X}}$, which we will always assume separable (that is, with a countable orthormal system).

Moreover, we denote p the canonical projection from $\mathcal{L}^2_{\rho_X}$ into $L^2_{\rho_X}$ such that $p : f \mapsto \tilde{f}$, with $\tilde{f} = \{g \in \mathcal{L}^2_{\rho_X}, \text{ s.t. } f \equiv g\}$. In the following, we will denote by f either the function $f \in \mathcal{L}^2_{\rho_X}$ or its class $\tilde{f} \in L^2_{\rho_X}$ when it does not induce any confusion: most of our assumptions, properties and proofs do not depend on the chosen representant function in the class. When $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, and $\rho_{\mathcal{X}}$ is a Borel-measure with full support, the difference between $\mathcal{L}^2_{\rho_X}$ and $L^2_{\rho_X}$ is traditionally neglected. When the measure does not have full support, as shown later in this section, the distinction is more important.

Assumption **(A2)** ensures that every function in $\mathcal{H}$ is square-integrable, that is, if $\mathbb{E}[K(X,X)] < \infty$, then $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$; for example, for $f = K_z$, $z \in \mathcal{X}$, $\|K_z\|^2_{L^2_{\rho_X}} = \mathbb{E}[K(X,z)^2] \leqslant K(z,z)\mathbb{E}K(X,X)$ (see more details in the Appendix A, Proposition 10).

The minimization problem will appear to be an approximation problem in $\mathcal{L}^2_{\rho_X}$, for which we will build estimates in $\mathcal{H}$. However, to derive theoretical results, it is easier to consider it as an approximation problem in the Hilbert space $L^2_{\rho_X}$, building estimates in $p(\mathcal{H})$, where $p(\mathcal{H})$ is the image of $\mathcal{H}$ via the mapping $i \circ p \ : \mathcal{H} \xrightarrow{i} \mathcal{L}^2_{\rho_X} \xrightarrow{p} L^2_{\rho_X}$, where i is the trivial injection and p the canonical projection.

Under a few additional often-made assumptions ($\mathcal{X}$ compact, ρ_X full support and K continuous), the spaces $\mathcal{H}$ and $p(\mathcal{H})$ may be identified, as the application $i \circ p$ will also be injective: in all the following propositions one may replace $p(\mathcal{H})$ by $\mathcal{H}$ to simplify understanding. However in the general setting there may exist functions $f \in \mathcal{H} \backslash \{0\}$ such that $\|f\|_{\mathcal{L}^2_{\rho_X}} = 0$ (thus $i \circ p$ non injective). This may for example occur if the support of ρ_X is strictly included in $\mathcal{X}$, and f is zero on this support, but not identically zero. See the Appendix A.5 for more details.

2.3. *Minimization problem.* We are interested in minimizing the following quantity, which is the *prediction error* of a function f, defined for any function in $\mathcal{L}^2_{\rho_X}$:

$$\varepsilon(f) = \frac{1}{2}\mathbb{E}\left[(f(X) - Y)^2\right]. \tag{2.1}$$

Since for $f \in \mathcal{L}^2_{\rho_X}$, $\varepsilon(f)$ only depends on the class $\tilde{f} \in L^2_{\rho_X}$, this also defines a functional ε on $L^2_{\rho_X}$.

We are looking for a function with a low prediction error in the particular function space $\mathcal{H}$, that is we aim to minimize $\varepsilon(f)$ over $f \in \mathcal{H}$. We have for $f \in L^2_{\rho_X}$:

$$\begin{aligned} \varepsilon(f) &= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \left\langle f, \int_{\mathcal{Y}} y d\rho_{Y|X=\cdot}(y) \right\rangle_{L^2_{\rho_X}} + \frac{1}{2}\mathbb{E}[Y^2] \\ &= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \langle f, \mathbb{E}\left[Y|X=\cdot\right]\rangle_{L^2_{\rho_X}} + \frac{1}{2}\mathbb{E}[Y^2]. \end{aligned} \tag{2.2}$$

A minimizer g of $\varepsilon(g)$ over $L^2_{\rho_X}$ is known to be such that $g(X) = \mathbb{E}[Y|X]$. Such a function is generally referred to as the regression function, and denoted g_ρ as it only depends on ρ. It is moreover unique (as an element of $L^2_{\rho_X}$). An important property of the prediction error is that:

$$\forall f \in L^2_{\rho_X}, \qquad \varepsilon(f) - \varepsilon(g_\rho) = \frac{1}{2}\|f - g_\rho\|^2_{L^2_{\rho_X}}. \tag{2.3}$$

It means that minimizing prediction error is equivalent to minimizing the distance in $L^2_{\rho_X}$ to the regression function. In this paper, we will not consider convergence in the norm $\|\cdot\|_{\mathcal{H}}$, because in general g_ρ does not belong to $\mathcal{H}$.

We are thus interested in approximating a function in $\mathcal{L}^2_{\rho_X}$ by functions in $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$. As a consequence, we need to consider closures with respect to $\|\cdot\|_{L^2_{\rho_X}}$. We traditionally denote the closure of any $F \subset L^2_{\rho_X}$ in $L^2_{\rho_X}$ as limits in $L^2_{\rho_X}$ of elements in F, that is:

$$\overline{F} = \Big\{ f \in L^2_{\rho_X} \ / \ \exists (f_n) \subset F, \|f_n - f\|_{L^2_{\rho_X}} \to 0 \Big\}.$$

Especially, for $p(\mathcal{H}) \subset L^2_{\rho_X}$:

$$\overline{p(\mathcal{H})} \ = \ \Big\{ f \in L^2_{\rho_X} \ / \ \exists (f_n) \subset p(\mathcal{H}), \|f_n - f\|_{L^2_{\rho_X}} \to 0 \Big\}.$$

The space $\overline{p(\mathcal{H})}$ is a closed and convex subset in $L^2_{\rho_X}$. We can thus define

$$g_{\mathcal{H}} = \arg \min_{f \in \overline{p(\mathcal{H})}} \frac{1}{2} \mathbb{E}\left[(f(X) - Y)^2 \right] = \arg \min_{f \in \overline{p(\mathcal{H})}} \frac{1}{2} \|f - g_\rho\|^2_{L^2_{\rho_X}}$$

as the orthogonal projection of g_ρ on $\overline{p(\mathcal{H})}$, using the existence of the projection on any closed convex set in a Hilbert space. This leads to the following proposition (see formal proof in the Appendix A.1, Proposition 11):

Proposition 1 (Definition of best approximation function). *Assume **(A1-2)**. The minimum of $\varepsilon(f)$ in $\overline{p(\mathcal{H})}$ is attained at a certain $g_{\mathcal{H}}$ (which is unique and well defined in $L^2_{\rho_X}$).*

It is important to notice that we do not assume that $g_{\mathcal{H}}$ is in $p(\mathcal{H})$, that is the minimum in $\mathcal{H}$ is in general not attained. In the following, we will also consider a pointwise defined version of $g_{\mathcal{H}}$ (by choosing any representant of the equivalence class), still denoted $g_{\mathcal{H}} \in \mathcal{L}^2_{\rho_X}$.

Estimation from n i.i.d. observations builds a sequence $(g_n)_{n \in \mathbb{N}}$ in $\mathcal{H}$. We will prove under suitable conditions that it satisfies weak consistency, that is g_n ends up predicting as well as $g_{\mathcal{H}}$:

$$\mathbb{E}\left[\varepsilon(g_n) - \varepsilon(g_{\mathcal{H}}) \right] \xrightarrow{n \to \infty} 0 \ \Leftrightarrow \ \|g_n - g_{\mathcal{H}}\|_{\mathcal{L}^2_{\rho_X}} \xrightarrow{n \to \infty} 0.$$

We can make the following remarks:

- We have proved existence and uniqueness of a minimizer over $\overline{p(\mathcal{H})} \subset L^2_{\rho_X}$. However approaching $g_{\mathcal{H}}$ in $L^2_{\rho_X}$ with some sequence built in $p(\mathcal{H})$ is strictly equivalent to approaching any function in $\mathcal{L}^2_{\rho_X}$ which is in the equivalence class $g_{\mathcal{H}}$. This is why we may accept confusion as it does not change anything in proofs or algorithms. However, for other tasks than for prediction from the random variable X, the space $\mathcal{H}$ and $p(\mathcal{H})$ are not equivalent.

– Seen as a function of $f \in \mathcal{H}$, our loss function ε is not coercive (i.e., not strongly convex), as our covariance operator[3] Σ has no minimal eigenvalue (the sequence of eigenvalues decreases to zero). As a consequence, even if $g_{\mathcal{H}} \in \mathcal{H}$, g_n may not converge to $g_{\mathcal{H}}$ in $\mathcal{H}$, and when $g_{\mathcal{H}} \notin \mathcal{H}$, we shall even have $\|g_n\|_{\mathcal{H}} \to \infty$.

2.4. *Covariance operator.* We now define the *covariance operator* for the space $\mathcal{H}$ and probability distribution ρ_X. The spectral properties of such an operator have appeared to be a key point to characterize speed of estimators [16, 4, 5].

We define a linear operator $\Sigma : \mathcal{H} \to \mathcal{H}$ through

$$\forall (f,g) \in \mathcal{H}^2, \quad \langle f, \Sigma g \rangle_{\mathcal{H}} = \mathbb{E}\left[f(X)g(X)\right] = \int_{\mathcal{X}} f(x)g(x) d\rho_X(x).$$

This operator is the *covariance operator* (defined on the Hilbert space $\mathcal{H}$). Using the reproducing property, we have[4]:

$$\Sigma = \mathbb{E}\left[K_X \otimes K_X\right],$$

where for any elements $g, h \in \mathcal{H}$, we denote by $g \otimes h$ the operator from $\mathcal{H}$ to $\mathcal{H}$ defined as:

$$\begin{aligned} g \otimes h \ : \mathcal{H} &\to \mathcal{H} \\ f &\mapsto \langle f, h \rangle_K \ g. \end{aligned}$$

In finite dimension, i.e., $\mathcal{H} = \mathbb{R}^d$, for $g, h \in \mathbb{R}^d$, $g \otimes h$ may be identified to a rank-one matrix, that is, $g \otimes h = gh^\top = \left((g_i h_j)_{1 \leqslant i,j \leqslant d}\right) \in \mathbb{R}^{d \times d}$ as for any f, $(gh^\top)f = g(h^\top f) = \langle h, f \rangle_{\mathcal{H}} g$. In other words, $g \otimes h$ is a linear operator, whose image is included in $\mathrm{Vect}(g)$, the linear space spanned by g. Thus in finite dimension, Σ is the (non-centered) covariance matrix.

We have defined the covariance operator on the Hilbert space $\mathcal{H}$. If $f \in \mathcal{H}$, we have for all $z \in \mathcal{X}$, using the reproducing property:

$$\mathbb{E}[f(X)K(X,z)] = \mathbb{E}[f(X)K_z(X)] = \langle K_z, \Sigma f \rangle_{\mathcal{H}} = (\Sigma f)(z),$$

which shows that the operator Σ may be extended to any square-integrable function $f \in \mathcal{L}^2_{\rho_X}$. In the following, we extend such an operator as an endomorphism $\mathcal{T}$ from $L^2_{\rho_X}$ to $\mathcal{L}^2_{\rho_X}$ and by projection as an endomorphism $T = p \circ \mathcal{T}$ from $L^2_{\rho_X}$ to $L^2_{\rho_X}$. Note that $\mathcal{T}$ is well defined as $\int_{\mathcal{X}} g(t)\ K_t \ d\rho_{\mathcal{X}}(t)$ does not depend on the function g chosen in the class of equivalence of g.

[3] See definition below.

[4] This expectation is formally defined as a Bochner expectation (an extension of Lebesgue integration theory to Banach spaces, see [19]) in $\mathcal{L}(\mathcal{H})$ the set of endomorphisms of $\mathcal{H}$.

DEFINITION 1 (Extended covariance operator). *Assume **(A1-2)**. We define the operator $\mathcal{T}$ as follows*[5]*:*

$$\begin{aligned} \mathcal{T} \ : \ L^2_{\rho_X} &\to \mathcal{L}^2_{\rho_X} \\ \tilde{g} &\mapsto \int_{\mathcal{X}} g(t)\ K_t\ d\rho_{\mathcal{X}}(t), \end{aligned}$$

so that for any $z \in \mathcal{X}$, $\mathcal{T}(g)(z) = \int_{\mathcal{X}} g(x)\ K(x,z)\ d\rho_{\mathcal{X}}(t) = \mathbb{E}[g(X)K(X,z)]$.

We give here some of the most important properties of $\mathcal{T}$ and $T = p \circ \mathcal{T}$. The operator T (which is an endomorphism of the separable Hilbert space $L^2_{\rho_X}$) may be reduced in some Hilbertian eigenbasis of $L^2_{\rho_X}$. It allows us to define the power of such an operator T^r, which will allow us to quantify the regularity of the function $g_{\mathcal{H}}$. See proof in Appendix A.2, Proposition 18.

PROPOSITION 2 (Eigen-decomposition of T). *Assume **(A1-2)**. T is a bounded self-adjoint semi-definite positive operator on $L^2_{\rho_X}$, which is trace-class. There exists a Hilbertian eigenbasis $(\phi_i)_{i\in I}$ of the orthogonal supplement S of the null space* $\mathrm{Ker}(T)$*, with summable strictly positive eigenvalues $(\mu_i)_{i\in I}$. That is:*

- *$\forall i \in I,\ T\phi_i = \mu_i\phi_i$, $(\mu_i)_{i\in I}$ strictly positive such that $\sum_{i\in I}\mu_i < \infty$.*
- *$L^2_{\rho_X} =$* $\mathrm{Ker}(T) \overset{\perp}{\oplus} S$*, that is, $L^2_{\rho_X}$ is the orthogonal direct sum of* $\mathrm{Ker}(T)$ *and S.*

When the space S has finite dimension, then I has finite cardinality, while in general I is countable. Moreover, the null space $\mathrm{Ker}(T)$ may be either reduced to $\{0\}$, finite-dimensional or infinite-dimensional.

The linear operator $\mathcal{T}$ happens to have an image included in $\mathcal{H}$, and the eigenbasis of T in $L^2_{\rho_X}$ may also be seen as eigenbasis of Σ in $\mathcal{H}$ (See proof in Appendix A.2, Proposition 17):

PROPOSITION 3 (Decomposition of Σ). *Assume **(A1-2)**. The image of $\mathcal{T}$ is included in $\mathcal{H}$: $Im(\mathcal{T}) \subset \mathcal{H}$, that is, for any $f \in L^2_{\rho_X}$, $\mathcal{T}f \in \mathcal{H}$. Moreover, for any $i \in I$, $\phi_i^H = \frac{1}{\mu_i}\mathcal{T}\phi_i \in \mathcal{H} \subset \mathcal{L}^2_{\rho_X}$ is a representant for the equivalence class ϕ_i, that is $p(\phi_i^H) = \phi_i$. Moreover $\mu_i^{1/2}\phi_i^H$ is an orthonormal eigen-system of the orthogonal supplement $\mathcal{S}$ of the null space $Ker(\Sigma)$. That is:*

- *$\forall i \in I,\ \Sigma\phi_i^H = \mu_i\phi_i^H$.*

[5]This expectation is formally defined as a Bochner expectation in $\mathcal{H}$.

– $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$.

We have two decompositions of $L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$ and $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$. The two orthogonal supplements S and $\mathscr{S}$ happen to be related through the mapping $\mathcal{T}^{1/2}$, which we now define.

More generally, we define all powers $\mathcal{T}^r$ (as operator from $L^2_{\rho_X}$ to $\mathcal{H}$) and powers T^r (as operator from $L^2_{\rho_X}$ to $L^2_{\rho_X}$). Note the different conditions regarding r.

Definition 2 (Powers of T). *We define, for any $r \geqslant 0$, $T^r : L^2_{\rho_X} \to L^2_{\rho_X}$, for any $h \in \mathrm{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through:*

$$T^r \left(h + \sum_{i\in I} a_i \phi_i \right) = \sum_{i\in I} a_i \mu_i^r \phi_i.$$

Moreover, for any $r > 0$, T^r may be defined as a bijection from S into $Im(\mathcal{T}^r)$. We may thus define its unique inverse:

$$T^{-r} \ : \ Im(T^r) \ \to \ S.$$

Definition 3 (Powers of $\mathcal{T}$). *We define, for any $r \geqslant 1/2$, $\mathcal{T}^r : L^2_{\rho_X} \to \mathcal{H}$, for any $h \in \mathrm{Ker}(T)$ and $(a_i)_{i\in I}$ such that $\sum_{i\in I} a_i^2 < \infty$, through:*

$$\mathcal{T}^r \left(h + \sum_{i\in I} a_i \phi_i \right) = \sum_{i\in I} a_i \mu_i^r \phi_i^H.$$

For $r \geqslant 1/2$, we clearly have that $T^r = p \circ \mathcal{T}^r$. In order to define $\mathcal{T}^r$, we need $r \geqslant 1/2$, because $(\mu_i^{1/2} \phi^H)$ is an orthonormal system of $\mathscr{S}$. Moreover, from the definition of $\mathcal{T}^{1/2}$, it clearly defines an isometry from S to $\mathscr{S}$. The next proposition shows that $p(\mathscr{S}) = p(\mathcal{H})$ and thus S and $p(\mathcal{H})$ are isomorphic. (See proof in Appendix A.2, Proposition 18):

Proposition 4 (Isometry between supplements). *$\mathcal{T}^{1/2} : S \to \mathscr{S}$ is an isometry. Moreover, $\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$ and $T^{1/2} : S \to p(\mathcal{H})$ is an isomorphism.*

This proposition may be understood as an extension of the following, which is a consequence of Mercer's theorem when we consider a compact set $\mathcal{X}$ and a continuous kernel on it ([4, 20]). The proposition above makes no topological assumptions about $\mathcal{X}$, the kernel K, and also applies to cases where $p(\mathcal{H})$ may be much smaller than $\mathcal{H}$.

PROPOSITION 5 (Isometry for Mercer kernels). *If $\mathcal{H}$ is a reproducing kernel Hilbert space associated with a Mercer kernel K (K continuous on the compact set $\mathcal{X}$), and* $\operatorname{supp}(\rho) = \mathcal{X}$, *then* $\mathcal{H} = T^{1/2}\left(L^2_{\rho_X}\right)$ *and* $T^{1/2} : S \to p(\mathcal{H})$ *is an isometrical isomorphism.*

In the general context, Proposition 4 has the following consequences:

- $T^{1/2}(S) = p(\mathcal{H})$, that is any element of $p(\mathcal{H})$ may be expressed as $T^{1/2}g$ for some $g \in L^2_{\rho_X}$.
- $\forall r \geqslant 1/2,\ T^r(S) \subset p(\mathcal{H})$, because for $r \geqslant 1/2$, $T^r(S) \subset T^{1/2}(S)$, that is, with large powers r, the image of T^r is in the projection of the Hilbert space.
- $\forall r > 0,\ \overline{T^r(L^2_{\rho_X})} = S = \overline{T^{1/2}(L^2_{\rho_X})} = \overline{p(\mathcal{H})}$, because (a) $T^{1/2}(L^2_{\rho_X}) = p(\mathcal{H})$ and (b) for any $r > 0$, $\overline{T^r(L^2_{\rho_X})} = S$. In other words, elements of $\overline{p(\mathcal{H})}$ (on which our minimization problem attains its minimum), may seen as limits (in $L^2_{\rho_X}$) of elements of $T^r(L^2_{\rho_X})$, for any $r > 0$.

In the following, the regularity of the function $g_{\mathcal{H}}$ will be characterized by the fact that $g_{\mathcal{H}}$ belongs to the space $T^r(L^2_{\rho_X})$ (and not only to its closure), for a specific $r > 0$ (see Section 2.6). The sequence of spaces $\{T^r(L^2_{\rho_X})\}_{r>0}$ is thus a decreasing (r increasing) sequence of subspaces of $L^2_{\rho_X}$ such that any of them is dense in $\overline{p(\mathcal{H})}$, and $T^r(L^2_{\rho_X}) \subset p(\mathcal{H})$ if and only if $r \geqslant 1/2$.

Finally, although we will not use it in the rest of the paper, we can state a version of Mercer's theorem, which does not make any more assumptions that are required for defining RKHSs.

PROPOSITION 6 (Kernel decomposition). *Assume* ***(A1-2)****. We have for all $x, y \in \mathcal{X}$,*

$$K(x,y) = \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y) + g(x,y),$$

and we have for all $x \in \mathcal{X}$, $\int_{\mathcal{X}} g(x,y)^2 d\rho_X(y) = 0$. Moreover, the convergence of the series is absolute.

We thus obtain a version of Mercer's theorem (see Appendix A.5.3) without any topological assumptions. Moreover, note that (a) $\mathscr{S}$ is also an RKHS, with kernel $(x,y) \mapsto \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$ and (b) that given the decomposition above, the optimization problem in $\mathscr{S}$ and $\mathcal{H}$ have equivalent solutions. Moreover, considering the algorithm below, the estimators we consider will almost surely build equivalent functions (see Appendix A.4). Thus, we could assume without loss of generality that the kernel K is exactly equal to its expansion $\sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$.

2.5. *Examples.* The property $\overline{p(\mathcal{H})} = S$, stated after Proposition 5, is important to understand what the space $\overline{p(\mathcal{H})}$ is, as we are minimizing over this closed and convex set. As a consequence the space $p(\mathcal{H})$ is dense in $L^2_{\rho_X}$ if and only if T is injective (or equivalently, $\mathrm{Ker}(L^2_{\rho_X}) = \{0\} \Leftrightarrow S = L^2_{\rho_X}$). We detail below a few classical situations in which different configurations for the "inclusion" $p(\mathcal{H}) \subset \overline{p(\mathcal{H})} \subset L^2_{\rho_X}$ appear:

1. **Finite-dimensional setting with linear kernel**: in finite dimension, with $\mathcal{X} = \mathbb{R}^d$ and $K(x,y) = x^\top y$, we have $\mathcal{H} = \mathbb{R}^d$, with the scalar product in $\langle u, v\rangle_H = \sum_{i=1}^d u_i v_i$. If the support of ρ_X has non-empty interior, then $\overline{p(\mathcal{H})} = p(\mathcal{H})$: $g_{\mathcal{H}}$ is the best linear estimator. Moreover, we have $p(\mathcal{H}) = \overline{p(\mathcal{H})} \subsetneq L^2_{\rho_X}$ in general. Moreover, $\mathrm{Ker}(\mathcal{T})$ is the set of functions such that $\mathbb{E}Xf(X) = 0$ (which is a large space), while $\mathrm{Ker}(\Sigma) = (\mathrm{span}\{\mathrm{supp}(\rho_X)\})^\perp$, where the orthogonal space is considered in $\mathbb{R}^d$. If ρ_X has a support for non-empty interior, it is reduced to $\{0\}$.
2. **Translation invariant kernels** for instance the Gaussian kernel over $\mathcal{X} = \mathbb{R}^d$, with X following a distribution with full support in $\mathbb{R}^d$: in such a situation we have $p(\mathcal{H}) \subsetneq \overline{p(\mathcal{H})} = L^2_{\rho_X}$. This last equality holds more generally for all universal kernels, which include all kernels of the form $K(x,y) = q(x-y)$ where q has a summable strictly positive Fourier transform [21, 22]. These kernels are exactly the kernels such that T is an injective endomorphism of $L^2_{\rho_X}$. In all these cases, the null spaces of $\mathcal{T}$ and Σ are equal to $\{0\}$.
3. **Splines over the circle:** When $X \sim \mathcal{U}[0;1]$ and $\mathcal{H}$ is the set of m-times periodic weakly differentiable functions (see Section 5), we have in general $p(\mathcal{H}) \subsetneq \overline{p(\mathcal{H})} \subsetneq L^2_{\rho_X}$. In such a case, $\ker(T) = \mathrm{span}(x \mapsto 1)$, and $\overline{p(\mathcal{H})} \oplus \mathrm{span}(x \mapsto 1) = L^2_{\rho_X}$, that is we can approximate any zero-mean function.

Many examples and more details may be found in [3, 23, 20]. In particular, kernels on non-vectorial objects may be defined (e.g., sequences, graphs or measures).

2.6. *Convergence rates.* In order to be able to establish rates of convergence in this infinite-dimensional setting, we have to make assumptions on the objective function and on the covariance operator eigenvalues. In order to account all cases (finite and infinite dimensions), we now consider eigenvalues ordered in non-increasing order, that is, we assume that the set I is either $\{0, \dots, d-1\}$ if the underlying space is d-dimensional or $\mathbb{N}$ if the underlying space has infinite dimension.

(A3) We denote $(\mu_i)_{i\in I}$ the sequence of non-zero eigenvalues of the operator T, in decreasing order. We assume $\mu_i \leqslant \frac{s^2}{i^\alpha}$ for some $\alpha > 1$ (so that $\mathrm{tr}(T) < \infty$), with $s \in \mathbb{R}_+$.

(A4) $g_{\mathcal{H}} \in T^r\left(\mathcal{L}^2_{\rho_X}\right)$ with $r \geqslant 0$, and as a consequence $\|T^{-r}(g_{\mathcal{H}})\|_{L^2_{\rho_X}} < \infty$.

We chose such assumptions in order to make the comparison with the existing literature as easy as possible, for example [8, 5]. However, some other assumptions may be found as in [24, 6].

Dependence on α and r. The two parameters r and α intuitively parameterize the strength of our assumptions:

- In assumption **(A3)** a bigger α makes the assumption stronger: it means the reproducing kernel Hilbert space is smaller, that is if **(A3)** holds with some constant α, then it also holds for any $\alpha' < \alpha$. Moreover, if T is reduced in the Hilbertian basis $(\phi_i)_i$ of $L^2_{\rho_X}$, we have an effective search space $\mathscr{S} = \{\sum_{i=1}^\infty a_i\phi_i^H / \sum_{i=1}^\infty \frac{a_i^2}{\mu_i} < \infty\}$: the smaller the eigenvalues, the smaller the space.
- In assumption **(A4)**, for a fixed α, a bigger r makes the assumption stronger, that is the function is actually smoother. Indeed, considering that **(A4)** may be rewritten $g_{\mathcal{H}} \in T^r(\mathcal{L}^2_{\rho_X})$ and for any $r < r'$, $T^{r'}(\mathcal{L}^2_{\rho_X}) \subset T^r(\mathcal{L}^2_{\rho_X})$. In other words, $\{T^r\left(\mathcal{L}^2_{\rho_X}\right)\}_{r\geq 0}$ are decreasing (r growing) subspaces of $\mathcal{L}^2_{\rho_X}$.
  For $r = 1/2$, $T^{1/2}(\mathcal{L}^2_{\rho_X}) = \mathcal{H}$; moreover, for $r \geqslant 1/2$, our best approximation function $g_{\mathcal{H}} \in \overline{p(\mathcal{H})}$ is in fact in $p(\mathcal{H})$, that is the optimization problem in the RKHS $\mathcal{H}$ is attained by a function of finite norm.

Related assumptions. The assumptions **(A3)** and **(A4)** are adapted to our theoretical results, but some stricter assumptions are often used, that make comparison with existing work more direct. For comparison purposes, we will also use:

(a3) For any $i \in I = \mathbb{N}$, $u^2 \leqslant i^\alpha\mu_i \leqslant s^2$ for some $\alpha > 1$ and $u, s \in \mathbb{R}_+$.

(a4) We assume the coordinates $(\nu_i)_{i\in\mathbb{N}}$ of $g_{\mathcal{H}} \in L^2_{\rho_X}$ in the eigenbasis $(\phi_i)_{i\in\mathbb{N}}$ (for $\|.\|_2$) of T are such that $\nu_i i^{\delta/2} \leqslant W$, for some $\delta > 1$ and $W \in \mathbb{R}_+$ (so that $\|g_{\mathcal{H}}\|_2 < \infty$).

Assumption **(a3)** directly imposes that the eigenvalues of T decay at rate $i^{-\alpha}$ (which imposes that there are infinitely many), and thus implies **(A3)**. Together, assumptions **(a3)** and **(a4)**, imply assumptions **(A3)** and **(A4)**,

with $\delta > 1 + 2\alpha r$. Indeed, we have

$$\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = \sum_{i \in \mathbb{N}} \nu_i^2 \mu_i^{-2r} \leqslant \frac{W^2}{u^{4r}} \sum_{i \in \mathbb{N}} i^{-\delta + 2\alpha r},$$

which is finite for $2\alpha r - \delta < -1$. Thus, the supremum element of the set of r such that **(A4)** holds is such that $\delta = 1 + 2\alpha r$. Thus, when comparing assumptions **(A3-4)** and **(a3-4)**, we will often make the identification above, that is, $\delta = 1 + 2\alpha r$.

The main advantage of the new assumptions is their interpretation when the basis $(\phi_i)_{i \in I}$ is common for several RKHSs (such as the Fourier basis for splines, see in Section 5): **(a4)** describes the decrease of the coordinates of the best function $g_{\mathcal{H}} \in L^2_{\rho_X}$ independently of the chosen RKHS. Thus, the parameter δ characterizes the prediction function, while the parameter α characterizes the RKHS.

3. Stochastic approximation in Hilbert spaces. In this section, we consider estimating a prediction function $g \in \mathcal{H}$ from observed data, and we make the following assumption:

(A5) For $n \geqslant 1$, the random variables $(x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ are independent and identically distributed with distribution ρ.

Our goal is to estimate a function $g \in \mathcal{H}$ from data, such that $\varepsilon(g) = \frac{1}{2}\mathbb{E}(Y - g(X))^2$ is as small as possible. As shown in Section 2, this is equivalent to minimizing $\|g - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$. The two main approaches to define an estimator is by regularization or by stochastic approximation (and combination thereof). See also approaches by early-stopped gradient descent on the empirical risk in [25].

3.1. *Regularization and linear systems.* Given n observations, regularized empirical risk minimization corresponds to minimizing with respect to $g \in \mathcal{H}$ the following objective function:

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \frac{\lambda}{2} \|g\|^2_{\mathcal{H}}.$$

Although the problem is formulated in a potentially infinite-dimensional Hilbert space, through the classical representer theorem [26, 3, 2], the unique (if $\lambda > 0$) optimal solution may be expressed as $\hat{g} = \sum_{i=1}^{n} a_i K_{x_i}$, and $a \in \mathbb{R}^n$ may be obtained by solving the linear system $(\mathbf{K} + \lambda I)a = \mathbf{y}$, where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix, a.k.a. the Gram matrix, composed

of pairwise kernel evaluations $\mathbf{K}_{ij} = K(x_i, x_j)$, $i, j = 1, \ldots, n$, and $\mathbf{y}$ is the n-dimensional vector of all n responses y_i, $i = 1, \ldots, n$.

The running-time complexity to obtain $a \in \mathbb{R}^n$ is typically $O(n^3)$ if no assumptions are made, but several algorithms may be used to lower the complexity and obtain an approximate solution, such as conjugate gradient [27] or column sampling (a.k.a. Nyström method) [28, 29, 30].

In terms of convergence rates, assumptions **(a3-4)** allow to obtain convergence rates that decompose $\varepsilon(\hat{g}) - \varepsilon(g_{\mathcal{H}}) = \frac{1}{2}\|\hat{g} - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$ as the sum of two asymptotic terms [5, 30, 24]:

- Variance term: $O(\sigma^2 n^{-1} \lambda^{-1/\alpha})$, which is decreasing with λ, where σ^2 characterizes the noise variance, for example, in the homoscedastic case (i.i.d. additive noise), the marginal variance of the noise; see assumption **(A6)** for the detailed assumption that we need in our stochastic approximation context.
- Bias term: $O(W^2 \lambda^{\min\{(\delta-1)/\alpha, 2\}})$, which is increasing with λ. Note that the corresponding r from assumptions **(A3-4)** is $r = (\delta - 1)/2\alpha$, and the bias term becomes proportional to $\lambda^{\min\{2r, 2\}}$.

There are then two regimes:

- Optimal predictions: If $r < 1$, then the optimal value of λ (that minimizes the sum of two terms and makes them equivalent) is proportional to $n^{-\alpha/(2r\alpha+1)} = n^{-\alpha/\delta}$ and the excess prediction error $\|\hat{g} - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O(n^{-2\alpha r/(2\alpha r+1)}) = O(n^{-1+1/\delta})$, and the resulting procedure is then "optimal" in terms of estimation of $g_{\mathcal{H}}$ in $L^2_{\rho_X}$ (see Section 4 for details).
- Saturation: If $r \geqslant 1$, where the optimal value of λ (that minimizes the sum of two terms and makes them equivalent) is proportional to $n^{-\alpha/(2\alpha+1)}$, and the excess prediction error is less than $= O(n^{-2\alpha/(2\alpha+1)})$, which is suboptimal. Although assumption **(A4)** is valid for a larger r, the rate is the same than if $r = 1$.

In this paper, we consider a stochastic approximation framework with improved running-time complexity and similar theoretical behavior than regularized empirical risk minimization.

3.2. *Stochastic approximation.* Using the reproducing property, we have for any $g \in \mathcal{H}$, $\varepsilon(g) = \frac{1}{2}\mathbb{E}(Y - g(X))^2 = \frac{1}{2}\mathbb{E}(Y - \langle g, K_X\rangle_{\mathcal{H}})^2$, with gradient (defined with respect to the dot-product in $\mathcal{H}$) $\varepsilon'(g) = -\mathbb{E}[(Y - \langle g, K_X\rangle_{\mathcal{H}})K_X]$.

Thus, for each pair of observations (x_n, y_n), we have $\varepsilon'(g) = -\mathbb{E}[(y_n -$

$\langle g, K_{x_n}\rangle_{\mathcal{H}})K_{x_n}\big]$, and thus, the quantity $\big[-(y_n - \langle g, K_{x_n}\rangle_{\mathcal{H}})K_{x_n}\big] = \big[-(y_n - g(x_n))\big)K_{x_n}\big]$ is an unbiased stochastic gradient. We thus consider the stochastic gradient recursion, in the Hilbert space $\mathcal{H}$, started from a function $g_0 \in \mathcal{H}$ (taken to be zero in the following):

$$g_n = g_{n-1} - \gamma_n\big[y_n - \langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}}\big]K_{x_n} = g_{n-1} - \gamma_n\big[y_n - g_{n-1}(x_n)\big]K_{x_n}.$$

We may also apply the recursion using representants. Indeed, if $g_0 = 0$, which we now assume, then for any $n \geqslant 1$,

$$g_n = \sum_{i=1}^{n} a_i K_{x_i},$$

with the following recursion on the sequence $(a_n)_{n\geqslant 1}$:

$$a_n = -\gamma_n(g_{n-1}(x_n) - y_n) = -\gamma_n\left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_i\right).$$

We also output the averaged iterate defined as

$$\overline{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} g_k = \frac{1}{n+1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{i} a_j\Big)K_{x_i}. \tag{3.1}$$

Running-time complexity. The running time complexity is $O(i)$ for iteration i—if we assume that kernel evaluations are $O(1)$, and thus $O(n^2)$ after n steps. Several authors have considered expanding g_n on a subset of all (K_{x_i}), which allows to bring down the complexity of each iteration and obtain an overall linear complexity is n [31, 32], but this comes at the expense of not obtaining the sharp generalization errors that we obtain in this paper. Note that when studying regularized least-squares problem (i.e., adding a penalisation term), one has to update every coefficient $(a_i)_{1\leqslant i\leqslant n}$ at step n. More content on learning with kernels may be found in [33], and references therein.

Relationship to previous works. Such type of algorithms have been studied before [8, 34, 33, 7, 35], under various forms. Especially, in [34, 9, 33, 35] a regularization term is added to the loss function (thus considering the following problem: $\arg\min_{f\in\mathcal{H}} \varepsilon(f) + \lambda||f||_K^2$). In [8, 7], neither regularization nor averaging procedure are considered, but in the second case, multiple pass through the data are considered. In [35], a non-regularized averaged procedure equivalent to ours is considered. However, the step-sizes γ_n which are proposed, as well as the corresponding analysis, are different. Our step-sizes are larger and our analysis uses more directly the underlying linear algebra to obtain better rates (while the proof of [35] is applicable to all smooth losses).

Learning rate. We are mainly interested in two different types of step sizes (a.k.a. learning rates): the sequence $(\gamma_i)_{1\leqslant i\leqslant n}$ may be either:

1. a subsequence of a universal sequence $(\gamma_i)_{i\in\mathbb{N}}$, we refer to this situation as the "online setting". Our bounds then hold for any of the iterates.
2. a sequence of the type $\gamma_i = \Gamma(n)$ for $i \leqslant n$, which will be referred to as the "finite horizon setting": in this situation the number of samples is assumed to be known and fixed and we chose a constant step size which may depend on this number. Our bound then hold only for the last iterate.

In practice it is important to have an online procedure, to be able to deal with huge amounts of data (potentially infinite). However, the analysis is easier in the "finite horizon" setting. Some doubling trick allows to pass to varying steps [36], but it is not satisfactory in practice as it creates jumps at every n which is a power of two.

3.3. *Extra regularity assumptions.* We denote by $\Xi = (Y - g_{\mathcal{H}}(X))K_X$ the residual, a random element of $\mathcal{H}$. We have $\mathbb{E}[\Xi] = 0$ but in general we do not have $\mathbb{E}[\Xi|X] = 0$ (unless the model of homoscedastic regression is well specified). We make the following extra assumption:

(A6) There exists $R > 0$ and $\sigma > 0$ such that $\mathbb{E}[\Xi \otimes \Xi] \preccurlyeq \sigma^2\Sigma$, and $\mathbb{E}(K(X,X)K_X \otimes K_X) \preccurlyeq R^2\Sigma$ where $\preccurlyeq$ denotes the order between self-adjoint operators.

In other words, for any $f \in \mathcal{H}$, we have: $\mathbb{E}\big[(Y-g_{\mathcal{H}}(X))^2 f(X)^2\big] \leqslant \sigma^2\mathbb{E}[f(X)^2]$ and $\mathbb{E}\big[K(X,X)f(X)^2\big] \leqslant R^2\mathbb{E}[f(X)^2]$.

In the well specified homoscedastic case, we have that $(Y-g_{\mathcal{H}}(X))$ is independent of X and with $\sigma^2 = \mathbb{E}\left[(Y - g_{\mathcal{H}}(X))^2\right]$, $\mathbb{E}[\Xi|X] = \sigma^2\Sigma$ is clear: the constant σ^2 in the first part of our assumption characterizes the noise amplitude. Moreover the second part assumption is clearly satisfied if $K(X,X)$ is almost surely bounded by R^2: this constant can then be understood as the radius of the set of our data points. However, our analysis holds in these more general set-ups where only fourth order moment of $\|K_x\|_{\mathcal{H}} = K(x,x)^{1/2}$ are finite.

We first present the results in the finite horizon setting in Section 3.4 before turning to online setting in Section 3.5.

3.4. *Main results (finite horizon).* We can first get some guarantee on the consistency of our estimator, for any small enough constant step-size:

THEOREM 1. *Assume **(A1-6)**, then for any constant choice $\gamma_n = \gamma_0 < \frac{1}{2R^2}$, the prediction error of $\bar{g}_n$ converges to the one of $g_{\mathcal{H}}$, that is:*

$$2\mathbb{E}\left[\varepsilon\left(\bar{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] = \mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \xrightarrow{n\to\infty} 0. \tag{3.2}$$

The expectation is considered with respect to the distribution of the sample $(x_i, y_i)_{1\leqslant i\leqslant n}$, as in all the following theorems (note that $\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}$ is itself a different expectation with respect to the law ρ_X).

Theorem 1 means that for the simplest choice of the learning rate as a constant, our estimator tends to the perform as well as the best estimator in the class $\mathcal{H}$. Note that in general, the convergence in $\mathcal{H}$ is meaningless if $r < 1/2$. The following results will state some assertions on the speed of such a convergence; our main result, in terms of generality is the following:

THEOREM 2 (Complete bound, γ constant, finite horizon). *Assume **(A1-6)** and $\gamma_i = \gamma = \Gamma(n)$, for $1 \leqslant i \leqslant n$. We have, with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$:*

$$\left(\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right)^{1/2} \leqslant \frac{\sigma/\sqrt{n}}{1-\sqrt{\gamma R^2}}\left(1 + C(\alpha)s^{\frac{2}{\alpha}}(\gamma n)^{\frac{1}{\alpha}}\right)^{\frac{1}{2}} + \frac{\|L_K^{-r} g_{\mathcal{H}}\|_{L^2_{\rho_X}}}{\gamma^r n^{\min\{r,1\}}}\left(1 + \frac{\sqrt{\gamma R^2}}{\sqrt{1-\gamma R^2}}\right).$$

For easier interpretation, we may derive a simple corollary:

COROLLARY 1 (γ constant, finite horizon). *Assume **A1-6**. If $\gamma R^2 \leqslant 1/4$, we have the simpler bound:*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant \frac{4\sigma^2}{n}\left(1 + \frac{1}{(\gamma_0 s^2)^{\frac{1}{\alpha}}}\right)^2 C(\alpha)(s^2\gamma n)^{\frac{1}{\alpha}} + 4\frac{\|L_K^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{2\min\{r,1\}}}. \tag{3.3}$$

We can make the following observations:

– **Proof**: Theorem 1 and Corollary 1 are directly derived from Theorem 2, which is proved in Appendix B.3: we derive for our algorithm a new error decomposition and bound the different sources of error via algebraic calculations. More precisely, following the proof in Euclidean space [13], we first analyze (in Appendix B.2) a closely related recursion (we replace $K_{x_n} \otimes K_{x_n}$ by its expectation Σ, and we thus refer to it as a semi-stochastic version of our algorithm):

$$g_n = g_{n-1} - \gamma_n(y_n K_{x_n} - \Sigma g_{n-1}).$$

It (a) leads to an easy computation of the main bias/variance terms of our result, (b) will be used to derive our main result by bounding the drifts between our algorithm and its semi-stochastic version.

– **Bias/variance interpretation**: The two main terms have a simple interpretation. The first one is a variance term, which shows the effect of the noise σ^2 on the error. It is bigger when σ gets bigger, and moreover it also gets bigger when γ is growing (bigger steps mean more variance). As for the second term, it is a bias term, which accounts for the distance of the initial choice (the null function in general) to the objective function. As a consequence, it is smaller when we make bigger steps.
– **Assumption (A4)**: Our assumption **(A4)** for $r > 1$ is stronger than **(A4)** for $r = 1$ but we do not improve the bound. Indeed the bias term (see comments below) cannot decrease faster than $O(n^{-2})$: this phenomenon in known as saturation [37]. To improve our results with $r > 1$ it may be interesting to consider another type of averaging. In the following, $r < 1$ shall be considered as the main and most interesting case.
– **Relationship to regularized empirical risk minimization**: Our bound ends up being very similar to bounds for regularized empirical risk minimization, with the identification $\lambda = \frac{1}{\gamma n}$. It is thus no surprise that once we optimize for the value of γ, we recover the same rates of convergence. Note that in order to obtain convergence, we require that γ is bounded, which corresponds to an equivalent λ which has to be lower-bounded by $1/n$.
– **Finite horizon**: Once again, this theorem holds in the finite horizon setting. That is we first choose the number of samples we are going to use, then the learning rate as a constant. It allows us to chose γ as a function of n, in order to balance the main terms in the error bound. The trade-off must be understood as follows: a bigger γ increases the effect of the noise, but a smaller one makes it harder to forget the initial condition.

We may now deduce the following corollaries, with specific optimized values of γ:

COROLLARY 2 (Optimal constant γ). *Assume **(A1-6)** and a constant step-size $\gamma_i = \gamma = \Gamma(n)$, for $1 \leqslant i \leqslant n$:*

1. *If* $\frac{\alpha-1}{2\alpha} < r$ *and* $\Gamma(n) = \gamma_0\, n^{\frac{-2\alpha\min\{r,1\}-1+\alpha}{2\alpha\min\{r,1\}+1}}$, $\gamma_0 R^2 \leqslant 1/4$, *we have:*

$$\mathbb{E}\left(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right) \leqslant A\; n^{-\frac{2\alpha\min\{r,1\}}{2\alpha\min\{r,1\}+1}}. \tag{3.4}$$

with $A = 4\left(1 + \frac{1}{(\gamma_0 s^2)^{\frac{1}{2\alpha}}}\right)^2 C(\alpha) s^{\frac{2}{\alpha}} \gamma_0^{1/\alpha}\sigma^2 + \frac{4}{\gamma_0^{2r}}||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}$.

2. *If* $0 < r < \frac{\alpha-1}{2\alpha}$, *with* $\Gamma(n) = \gamma_0$ *is constant, we have:*

$$\mathbb{E}\left(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right) \leqslant A\; n^{-\frac{\alpha-1}{2\alpha}}, \tag{3.5}$$

with the same constant A.

We can make the following observations:

- **Limit conditions:** Assumption **(A4)**, gives us some kind of "position" of the objective function with respect to our reproducing kernel Hilbert space. If $r \geqslant 1/2$ then $g_{\mathcal{H}} \in \mathcal{H}$. That means the regression function truly lies in the space in which we are looking for an approximation. However, it is not necessary neither to get the convergence result, which stands for any $r > 0$, nor to get the optimal rate (see definition in Section 4.2), which is also true for $\frac{\alpha-1}{2\alpha} < r < 1$.
- **Evolution with r and α:** As it has been noticed above, a bigger α or r would be a stronger assumption. It is thus natural to get a rate which improves with a bigger α or r: the function $(\alpha, r) \mapsto \frac{2\alpha r}{2\alpha r+1}$ is increasing in both parameters.
- **Different regions:** in Figure 1a, we plot in the plan of coordinates α, δ (with $\delta = 2\alpha r + 1$) our limit conditions concerning our assumptions, that is, $r = 1 \Leftrightarrow \delta = 2\alpha + 1$ and $\frac{\alpha-1}{2\alpha} = r \Leftrightarrow \alpha = \delta$. The region between the two green lines is the region for which the optimal rate of estimation is reached. The magenta dashed lines stands for $r = 1/2$, which has appeared to be meaningless in our context.
  The region $\alpha \geqslant \delta \Leftrightarrow \frac{\alpha-1}{2\alpha} > r$ corresponds to a situation where regularized empirical risk minimization would still be optimal, but with a regularization parameter λ that decays faster than $1/n$, and thus, our corresponding step-size $\gamma = 1/(n\lambda)$ would not be bounded as a function of n. We thus saturate our step-size to a constant and the generalization error is dominated by the bias term.
  The region $\alpha \leqslant (\delta - 1)/2 \Leftrightarrow r > 1$ corresponds to a situation where regularized empirical risk minimization reaches a saturating behaviour. In our stochastic approximation context, the variance term dominates.

3.5. *Theorem (online).* We now consider the second case when the sequence of step sizes does not depend on the number of samples we want to use (online setting).

The computation are more tedious in such a situation so that we will only state asymptotic theorems in order to understand the similarities and differences between the finite horizon setting and the online setting, especially in terms of limit conditions.

THEOREM 3 (Complete bound, $(\gamma_n)_n$ online). *Assume **(A1-6)**, assume for any i, $\gamma_i = \frac{\gamma_0}{i^\zeta}$.*

- *If $0 < r(1-\zeta) < 1$, if $0 < \zeta < \frac{1}{2}$ then*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant O\left(\frac{\sigma^2 (s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right) + O\left(\frac{||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{3.6}$$

- *If $0 < r(1-\zeta) < 1$, $\frac{1}{2} < \zeta$*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant O\left(\frac{\sigma^2 (s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} \frac{1}{n\gamma_n^2}\right) + O\left(\frac{||L_K^{-r} g_{\mathcal{H}}||^2_{L^2_{\rho_X}}}{(n\gamma_n)^{2r}}\right). \tag{3.7}$$

The constant in the $O(\cdot)$ notations only depend on γ_0 and α.

Theorem 3 is proved in Appendix B.4. In the first case, the main bias and variance terms are the same as in the finite horizon setting, and so is the optimal choice of ζ. However in the second case, the variance term behaviour changes: it does not decrease any more when ζ increases beyond $1/2$. Indeed, in such a case our constant averaging procedure puts to much weight on the first iterates, thus we do not improve the variance bound by making the learning rate decrease faster. Other type of averaging, as proposed for example in [38], could help to improve the bound.

Moreover, this extra condition thus changes a bit the regions where we get the optimal rate (see Figure 1b), and we have the following corollary:

COROLLARY 3 (Optimal decreasing γ_n). *Assume **(A1-6)** (in this corollary, $O(\cdot)$ stands for a constant depending on $\alpha, ||L_K^{-r} g_{\mathcal{H}}||_{L^2_{\rho_X}}, s, \sigma^2, \gamma_0$ and universal constants):*

1. *If $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$, with $\gamma_n = \gamma_0 n^{\frac{-2\alpha r - 1 + \alpha}{2\alpha r + 1}}$ for any $n \geqslant 1$ we get the rate:*

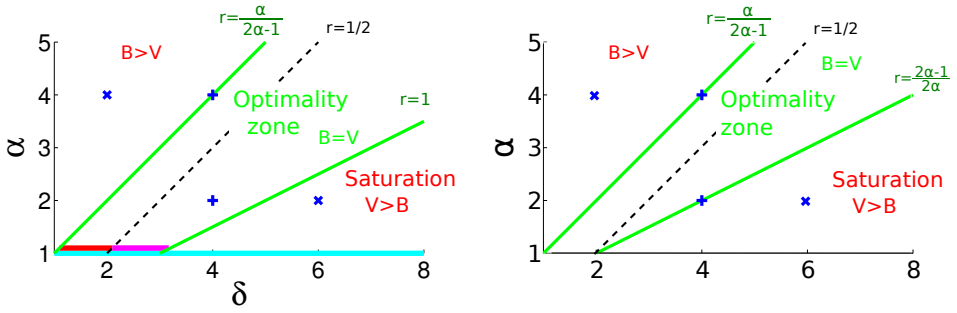
$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(n^{-\frac{2\alpha r}{2\alpha r+1}}\right). \tag{3.8}$$

2. *If* $\frac{2\alpha-1}{2\alpha} < r$, *with* $\gamma_n = \gamma_0 n^{-1/2}$ *for any* $n \geqslant 1$, *we get the rate:*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} = O\left(n^{-\frac{2\alpha-1}{2\alpha}}\right). \tag{3.9}$$

3. *If* $0 < r < \frac{\alpha-1}{2\alpha}$, *with* $\gamma_n = \gamma_0$ *for any* $n \geqslant 1$, *we get the rate given in* (3.5). *Indeed the choice of a constant learning rate naturally results in an online procedure.*

This corollary is directly derived from Theorem 3, balancing the two main terms. The only difference with the finite horizon setting is the shrinkage of the optimality region as the condition $r < 1$ is replaced by $r < \frac{2\alpha-1}{2\alpha} < 1$ (see Figure 1b). In the next section, we relate our results to existing work.



(a) Finite Horizon (b) Online

Figure 1: Behaviour of convergence rates: (left) finite horizon and (right) online setting. We describe in the (α, δ) plan (with $\delta = 2\alpha r + 1$) the different optimality regions : between the two green lines, we achieve the optimal rate. On the left plot the red (respectively magenta and cyan) lines are the regions for which Zhang (respectively Yao&Tarres and Ying&Pontil) proved to achieve the overall optimal rate (which may only be the case if $\alpha = 1$). The four blue points match the coordinates of the four couples (α, δ) that will be used in our simulations : they are spread over the different optimality regions.

4. Links with existing results. In this section, we relate our results from the previous section to existing results.

4.1. *Euclidean spaces.* Recently, Bach and Moulines showed in [13] that for least squares regression, averaged stochastic gradient descent achieved

a rate of $O(1/n)$, in a finite-dimensional Hilbert space (Euclidean space), under the same assumptions as above (except the first one of course), which is replaced by:

(A1') $\mathcal{H}$ is a d-dimensional Euclidean space.

They showed the following result:

PROPOSITION 7 (Finite-dimensions [13]). *Assume **(A1'), (A2-6)**. Then for any constant step-size $\gamma < \frac{1}{R^2}$, we have*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \frac{1}{2n}\left[\frac{\sigma\sqrt{d}}{1-\sqrt{\gamma R^2}} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\frac{1}{\sqrt{\gamma R^2}}\right]^2. \tag{4.1}$$

Thus with $\gamma = \frac{1}{4R^2}$ we get $\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \frac{2}{n}\left[\sigma\sqrt{d} + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\right]^2$.

We show that we can deduce such a theorem from Theorem 2 (and even with comparable constants). Indeed under **(A1')** we have:

- If $\mathbb{E}\left[||x_n||^2\right] \leqslant R^2$ then $\Sigma \preccurlyeq R^2 I$ and **(A3)** is true for any $\alpha \geq 1$ with $s^2 = R^2 d^\alpha$. Indeed $\lambda_i \leqslant R^2$ if $i \leqslant d$ and $\lambda_i = 0$ if $i > d+1$ so that for any $\alpha > 1, i \in \mathbb{N}^*$, $\lambda_i \leqslant R^2 \frac{d^\alpha}{i^\alpha}$.
- As we are in a finite-dimensional space **(A4)** is true for $r = 1/2$ as $||T^{-1/2} g_{\mathcal{H}}||^2_{\mathcal{L}^2_{\rho_X}} = ||g_{\mathcal{H}}||^2_{\mathcal{H}}$.

Under such remarks, the following corollary may be deduced from Theorem 2:

COROLLARY 4. *Assume **(A1'), (A2-6)**, then for any $\alpha > 1$, with $\gamma R^2 \leqslant 1/4$:*

$$\mathbb{E}\|\overline{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant \frac{4\sigma^2}{n}\left(1 + \frac{1}{(\gamma_0 R^2 d^\alpha)^{\frac{1}{2\alpha}}}\right)^2 C(\alpha)(R^2\gamma d^\alpha n)^{\frac{1}{\alpha}} + 4\frac{\|g_{\mathcal{H}}\|^2_{\mathcal{H}}}{n\gamma}.$$

So that, when $\alpha \to \infty$, $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)} \to_{\alpha\to\infty} 1$, and

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \frac{2}{n}\left(\sigma\sqrt{d}\left(1 + \frac{1}{\sqrt{d}}\right) + R\|g_{\mathcal{H}}\|_{\mathcal{H}}\frac{1}{\sqrt{\gamma R^2}}\right)^2.$$

This bound is easily comparable to (4.1) and shows that our analysis has not lost too much. Moreover our learning rate is proportional to $n^{\frac{-1}{2\alpha+1}}$ with

$r = 1/2$, so tends to behave like a constant when $\alpha \to \infty$, which recovers the constant step set-up from [13].

Moreover, N. Flammarion proved[6], using same kind of techniques, that their bound could be extended to:

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant 4\frac{\sigma^2 d}{n} + 4R^4 \frac{\|\Sigma^{-1/2} g_{\mathcal{H}}\|^2}{(\gamma R^2)^2 n^2}, \tag{4.2}$$

a result that may be deduced of the following more general corollary of our Theorem 2:

COROLLARY 5. *Assume* ***(A1')***, ***(A2-6)***, *and, for some* $q \geqslant 0$, $||\Sigma^{-q} g_{\mathcal{H}}||^2_{\mathcal{H}} = ||\Sigma^{-(q+1/2)} g_{\mathcal{H}}||^2_{L^2_{\rho_X}} < \infty$, *then:*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_*)\right] \leqslant 8\frac{\sigma^2 d}{n} + 4R^{4(q+1/2)} \frac{||\Sigma^{-q} g_*||^2_{\mathcal{H}}}{(n\gamma R^2)^{2(q+1/2)}}.$$

Such a result is derived from Theorem 2 with $\alpha \to \infty$, with $r = q + 1/2$. It bridges the gap between Proposition 7 ($q = 0$), and its extension (4.2) ($q = 1/2$). The constants 4 and 8 come from the upper bounds $(a+b)^2 \leqslant a^2 + b^2$ and $1 + 1/\sqrt{d} \leqslant 2$ and are thus non optimal.

Remark: linking our work to the finite-dimensional setting is made using the fact that our assumption **(A3)** is true for any $\alpha > 1$ and the fact that $C(\alpha) \to_{\alpha \to \infty} 1$.

4.2. *Optimal rates of estimation.* In some situations, our stochastic approximation framework leads to "optimal" rates of prediction in the following sense. In [5, Theorem 2] a minimax lower bound was given: let $\mathcal{P}(\alpha, r)$ $(\alpha > 1, r \in [1/2, 1])$ be the set of all probability measures ρ on $\mathcal{X} \times \mathcal{Y}$, such that:

- a.s., $|y| \leqslant M_\rho$,
- $T^{-r} g_\rho \in \mathcal{L}^2_{\rho(X)}$,
- the eigenvalues $(\mu_j)_{j \in \mathbb{N}}$ arranged in a non increasing order, are subject to the decay $\mu_j = O(j^{-\alpha})$.

Then the following minimax lower rate stands:

$$\liminf_{n \to \infty} \inf_{g_n} \sup_{\rho \in \mathcal{P}(b,r)} \mathbb{P}\left\{\varepsilon(g_n) - \varepsilon(g_\rho) > C n^{-2r\alpha/(2r\alpha+1)}\right\} = 1,$$

[6]Personnal communication, 05/2014

for some constant $C > 0$ where the infimum in the middle is taken over all algorithms as a map $((x_i, y_i)_{1 \leqslant i \leqslant n}) \mapsto g_n \in \mathcal{H}$.

When making assumptions **(a3-4)**, the assumptions regarding the prediction problem (i.e., the optimal function g_ρ) are summarized in the decay of the components of g_ρ in an orthonormal basis, characterized by the constant δ. Here, the minimax rate of estimation (see, e.g., [39]) is $O(n^{-1+1/\delta})$ which is the same as $O(n^{-2r\alpha/(2r\alpha+1)})$ with the identification $\delta = 2\alpha r + 1$.

That means the rate we get is optimal for $\frac{\alpha-1}{2\alpha} < r < 1$ in the finite horizon setting, and for $\frac{\alpha-1}{2\alpha} < r < \frac{2\alpha-1}{2\alpha}$ in the online setting. This is the region between the two green lines on Figure 1.

4.3. *Regularized stochastic approximation.* It is interesting to link our results to what has been done in [34] and [9] in the case of regularized least-mean-squares, so that the recursion is written:

$$g_n \quad = \quad g_{n-1} - \gamma_n \left((g_{n-1}(x_n) - y_n) K_{x_n} + \lambda_n g_{n-1} \right)$$

with $(g_{n-1}(x_n) - y_n) K_{x_n} + \lambda_n g_{n-1}$ an unbiased gradient of $\frac{1}{2}\mathbb{E}_\rho \left[(g(x) - y)^2 \right] + \frac{\lambda_n}{2} ||g||^2$. In [9] the following result is proved (*Remark 2.8* following *Theorem C*):

THEOREM 4 (Regularized, non averaged stochastic gradient[9]). *Assume that $T^{-r} g_\rho \in L^2_{\rho_X}$ for some $r \in [1/2, 1]$. Assume the kernel is bounded and $\mathcal{Y}$ compact. Then with probability at least $1 - \kappa$, for all $t \in \mathbb{N}$,*

$$\varepsilon(g_n) - \varepsilon(g_\rho) \leqslant O_\kappa \left(n^{-2r/(2r+1)} \right).$$

Where O_κ stands for a constant which depends on κ.

No assumption is made on the covariance operator beyond being trace class, but only on $\|T^{-r} g_\rho\|_{L^2_{\rho_X}}$ (thus no assumption **(A3)**). A few remarks may be made:

1. They get almost-sure convergence, when we only get convergence in expectation. We could perhaps derive a.s. convergence by considering moment bounds in order to be able to derive convergence in high probability and to use Borel-Cantelli lemma.
2. They only assume $\frac{1}{2} \leqslant r \leqslant 1$, which means that they assume the regression function to lie in the RKHS.

4.4. *Unregularized stochastic approximation.* In [8], Ying and Pontil studied the same unregularized problem as we consider, under assumption **(A4)**. They obtain the same rates as above ($n^{-2r/(2r+1)}\log(n)$) in both online case (with $0 \leqslant r \leqslant \frac{1}{2}$) and finite horizon setting ($0 < r$).

They led as an open problem to improve bounds with some additional information on some decay of the eigenvalues of T, a question which is answered here.

Moreover, Zhang [35] also studies stochastic gradient descent algorithms in an unregularized setting, also with averaging. As described in [8], his result is stated in the linear kernel setting but may be extended to kernels satisfying $\sup_{x\in\mathcal{X}} K(x,x) \leqslant R^2$. Ying and Pontil derive from Theorem 5.2 in [35] the following proposition:

PROPOSITION 8 (Short step-sizes [35]). *Assume we consider the algorithm defined in Section 3.2 and output $\bar{g}_n$ defined by equation (3.1). Assume the kernel K satisfies $\sup_{x\in\mathcal{X}} K(x,x) \leqslant R^2$. Finally assume g_ρ satisfies assumption **(A4)** with $0 < r < 1/2$. Then in the finite horizon setting, with $\Gamma(n) = \frac{1}{4R^2} n^{-\frac{2r}{2r+1}}$, we have:*

$$\mathbb{E}\left[\varepsilon\left(\bar{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] = O\left(n^{-\frac{2r}{2r+1}}\right).$$

Moreover, note that we may derive their result from Corollary 2. Indeed, using $\Gamma(n) = \gamma_0 n^{\frac{-2r}{2r+1}}$, we get a bias term which is of order $n^{\frac{-2r}{2r+1}}$ and a variance term of order $n^{-1+\frac{1}{2r\alpha+\alpha}}$ which is smaller. Our analysis thus recovers their convergence rate with their step-size. Note that this step-size they is significantly smaller than ours, and that the resulting bound is worse (but their result holds in more general settings than least-squares). See more details in Section 4.5.

4.5. *Summary of results.* All three algorithms are variants of the following:

$$\begin{aligned} g_0 &= 0 \\ \forall n \geqslant 1, \quad g_n &= (1-\lambda_n) g_{n-1} - \gamma_n (y_n - g_{n-1}(x_n)) K_{x_n}. \end{aligned}$$

But they are studied under different settings, concerning regularization, averaging, assumptions: we sum up in Table 1 the settings of each of these studies. For each of them, we consider the finite horizon settings, where results are generally better.

Algorithm type	Ass. **(A3)**	Ass. **(A4)**	γ_n	λ_n	Rate	Conditions
This paper (1)	yes	yes	1	0	$n^{-\frac{2\alpha}{2\alpha-1}}$	$r < \frac{\alpha-1}{2\alpha}$
This paper (2)	yes	yes	$n^{-\frac{2\alpha r+1-\alpha}{2\alpha r+1}}$	0	$n^{\frac{-2\alpha r}{2\alpha r+1}}$	$\frac{\alpha-1}{2\alpha} < r < 1$
This paper (3)	yes	yes	$n^{-\frac{\alpha+1}{2\alpha+1}}$	0	$n^{\frac{-2\alpha}{2\alpha+1}}$	$r > 1$
Zhang [35]	no	yes	$n^{\frac{-2r}{2r+1}}$	0	$n^{\frac{-2r}{2r+1}}$	$0 \leqslant r \leqslant \frac{1}{2}$
Tarrès & Yao [9]	no	yes	$n^{\frac{-2r}{2r+1}}$	$n^{\frac{-1}{2r+1}}$	$n^{\frac{-2r}{2r+1}}$	$\frac{1}{2} \leqslant r \leqslant 1$
Ying & Pontil [8]	no	yes	$n^{\frac{-2r}{2r+1}}$	0	$n^{\frac{-2r}{2r+1}}$	$r > 0$

TABLE 1

Summary of assumptions and results (step-sizes, rates and conditions) for our three regions of convergence and related approaches. We focus on finite-horizon results.

We can make the following observations:

- **Dependence of the convergence rate on** α: For learning in any kernel with $\alpha > 1$ we strictly improve the asymptotic rate compared to related methods that only assume summability of eigenvalues: indeed, the function $x \mapsto x/(x+1)$ is increasing on $\mathbb{R}^+$. If we consider a given optimal prediction function and a given kernel in which we are going to learn the function, considering the decrease in eigenvalues allows to adapt the step-size and obtain an improved learning rate. Namely, we improved the previous rate $\frac{-2r}{2r+1}$ up to $\frac{-2\alpha r}{2\alpha r+1}$.
- **Worst case result in** r: in the setting of assumptions **(a3,4)**, given δ, the optimal rate of convergence is known to be $O(n^{-1+1/\delta})$, where $\delta = 2\alpha r + 1$. We thus get the optimal rate, as soon as $\alpha < \delta < 2\alpha + 1$, while the other algorithms get the suboptimal rate $n^{\frac{\delta-1}{\delta+\alpha-1}}$ under various conditions. Note that this sub-optimal rate becomes close to the optimal rate when α is close to one, that is, in the worst case situation. Thus, in the worst-case (α arbitrarily close to one), all methods behave similarly, but for any particular instance where $\alpha > 1$, our rates are better.
- **Choice of kernel**: in the setting of assumptions **(a3,4)**, given δ, in order to get the optimal rate, we may choose the kernel (i.e., α) such that $\alpha < \delta < 2\alpha + 1$ (that is neither too big, nor too small), while other methods need to choose a kernel for which α is as close to one as possible, which may not be possible in practice.
- **Improved bounds**: Ying and Pontil [8] only give asymptotic bounds, while we have exact constants for the finite horizon case. Moreover there are some logarithmic terms in [8] which disappear in our analysis.
- **Saturation**: our method does saturate for $r > 1$, while the non-

averaged framework of [8] does not (but does not depend on the value of α). We conjecture that a proper non-uniform averaging scheme (that puts more weight on the latest iterates), we should get the best of both worlds.

5. Experiments on artificial data. Following [8], we consider synthetic examples with smoothing splines on the circle, where our assumptions **(A3-4)** are easily satisfied.

5.1. *Splines on the circle.* The simplest example to match our assumptions may be found in [1]. We consider $Y = g_\rho(X) + \varepsilon$, with $X \sim \mathcal{U}[\,0;1]$ is a uniform random variable in $[0,1]$, and g_ρ in a particular RKHS (which is actually a Sobolev space).

Let $\mathcal{H}$ be the collection of all zero-mean periodic functions on $[0;1]$ of the form

$$f : t \mapsto \sqrt{2}\sum_{i=1}^{\infty} a_i(f)\cos(2\pi i t) + \sqrt{2}\sum_{i=1}^{\infty} b_i(f)\sin(2\pi i t),$$

with

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty}(a_i(f)^2 + b_i(f)^2)(2\pi i)^{2m} < \infty.$$

This means that the m-th derivative of f, $f^{(m)}$ is in $\mathcal{L}^2([0\,;1])$. We consider the inner product:

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty}(2\pi i)^{2m}\left(a_i(f)a_i(g) + b_i(f)b_i(g)\right).$$

It is known that $\mathcal{H}$ is an RKHS and that the reproducing kernel $R_m(s,t)$ for $\mathcal{H}$ is

$$\begin{aligned} R_m(s,t) &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}[\cos(2\pi i s)\cos(2\pi i t) + \sin(2\pi i s)\sin(2\pi i t)] \\ &= \sum_{i=1}^{\infty}\frac{2}{(2\pi i)^{2m}}\cos(2\pi i(s-t)). \end{aligned}$$

Moreover the study of Bernoulli polynomials gives a close formula for $R(s,t)$, that is:

$$R_m(s,t) = \frac{(-1)^{m-1}}{(2m)!}B_{2m}\left(\{s-t\}\right),$$

with B_m denoting the m-th Bernoulli polynomial and $\{s-t\}$ the fractional part of $s-t$ [1].

We can derive the following proposition for the covariance operator which means that our assumption **(A3)** is satisfied for our algorithm in $\mathcal{H}$ when $X \sim \mathcal{U}[0;\,1]$, with $\alpha = 2m$, and $s = 2(1/2\pi)^m$.

PROPOSITION 9 (Covariance operator for smoothing splines). *If* $X \sim \mathcal{U}[0;\,1]$, *then in* $\mathcal{H}$*:*

1. *the eigenvalues of* Σ *are all of multiplicity 2 and are* $\lambda_i = (2\pi i)^{-2m}$,
2. *the eigenfunctions are* $\phi_i^c : t \mapsto \sqrt{2}\cos(2\pi i t)$ *and* $\phi_i^s : t \mapsto \sqrt{2}\sin(2\pi i t)$.

PROOF. For ϕ_i^c we have (a similar argument holds for ϕ_i^s):

$$\begin{aligned}\mathcal{T}(\phi_i^c)(s) &= \int_0^1 R(s,t)\sqrt{2}\cos(2\pi i t)dt \\ &= \left(\int_0^1 \frac{2}{(2i\pi)^{2m}}\sqrt{2}\cos(2\pi i t)^2 dt\right)\cos(2\pi i s) = \lambda_i\sqrt{2}\cos(2\pi i s) \\ &= \lambda_i\phi_i^c(s).\end{aligned}$$

It is well known that $(\phi_i^c, \phi_i^s)_{i\geqslant 0}$ is an orthonormal system (the Fourier basis) of the functions in $\mathcal{L}^2([\,0;1])$ with zero mean, and it is easy to check that $((2i\pi)^{-m}\phi_i^c, (2i\pi)^{-m}\phi_i^s)_{i\geqslant 1}$ is an orthonormal basis of our RKHS $\mathcal{H}$ (this may also be seen as a consequence of the fact that $\mathcal{T}^{1/2}$ is an isometry). □

Finally, considering $g_\rho(x) = B_{\delta/2}(x)$ with $\delta = 2\alpha r + 1 \in 2\mathbb{N}$, our assumption **(A4)** holds. Indeed it implies **(a3-4)**, with $\alpha > 1, \delta = 2\alpha r + 1$, since for any $k \in \mathbb{N}$, $B_k(x) = -2k!\sum_{i=1}^{\infty}\frac{\cos\left(2i\pi x - \frac{k\pi}{2}\right)}{(2i\pi)^k}$ (see, e.g., [40]).

We may notice a few points:

1. Here the eigenvectors do not depend on the kernel choice, only the renormalisation constant depends on the choice of the kernel. Especially the eigenbasis of $\mathcal{T}$ in $L^2_{\rho_X}$ doesn't depend on m. That can be linked with the previous remarks made in Section 4.
2. Assumption **(A3)** defines here the size of the RKHS: the smaller $\alpha = 2m$ is, the bigger the space is, the harder it is to learn a function.

In the next section, we illustrate on such a toy model our main results and compare our learning algorithm to Ying and Pontil's [8], Tarrès and Yao's [9] and Zhang's [35] algorithms.

5.2. *Experimental set-up.* We use $g_\rho(x) = B_{\delta/2}(x)$ with $\delta = 2\alpha r + 1$, as proposed above, with $B_1(x) = x - \frac{1}{2}$, $B_2(x) = x^2 - x + \frac{1}{6}$ and $B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x$.

We give in Figure 2 the functions used for simulations in a few cases that span our three regions. We also remind the choice of γ proposed for the 4 algorithms. We always use the finite horizon setting.

r	α	δ	K	g_ρ	$\frac{\log(\gamma)}{\log(n)}$ (this paper)	$\frac{\log(\gamma)}{\log(n)}$ (previous)
0.75	2	4	R_1	B_2	$-1/2 = -0.5$	$-3/5 = -0.6$
0.375	4	4	R_2	B_2	0	$-3/7 \simeq -0.43$
1.25	2	6	R_1	B_3	$-3/7 \simeq -0.43$	$-5/7 \simeq -0.71$
0.125	4	2	R_2	B_1	0	$-1/5 = -0.2$

TABLE 2

Different choices of the parameters α, r and the corresponding convergence rates and step-sizes. The (α, δ) coordinates of the four choices of couple "(kernel, objective function)" are mapped on Figure 1. They are spread over the different optimality regions.

5.3. *Optimal learning rate for our algorithm.* In this section, we empirically search for the best choice of a finite horizon learning rate, in order to check if it matches our prediction. For a certain number of values for n, distributed exponentially between 1 and $10^{3.5}$, we look for the best choice $\Gamma_{\text{best}}(n)$ of a constant learning rate for our algorithm up to horizon n. In order to do that, for a large number of constants $C_1, \cdots, C_p$, we estimate the expectation of error $\mathbb{E}[\varepsilon(\overline{g}_n(\gamma = C_i)) - \varepsilon(g_\rho)]$ by averaging over 30 independent sample of size n, then report the constant giving minimal error as a function of n in Figure 2. We consider here the situation $\alpha = 2, r = 0.75$. We plot results in a logarithmic scale, and evaluate the asymptotic decrease of $\Gamma_{\text{best}}(n)$ by fitting an affine approximation to the second half of the curve. We get a slope of -0.51, which matches our choice of -0.5 from Corollary 2. Although, our theoretical results are only upper-bounds, we conjecture that our proof technique also leads to lower-bounds in situations where assumptions (**a3-4**) hold (like in this experiment).

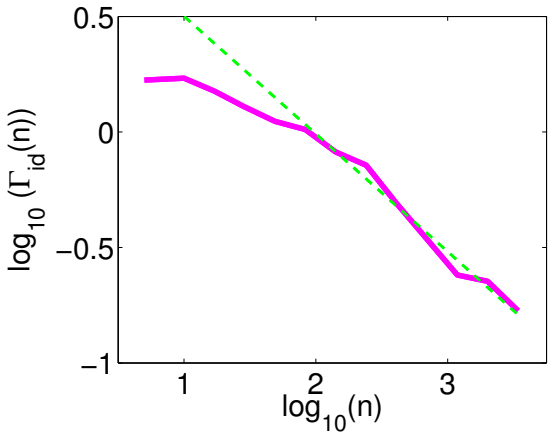
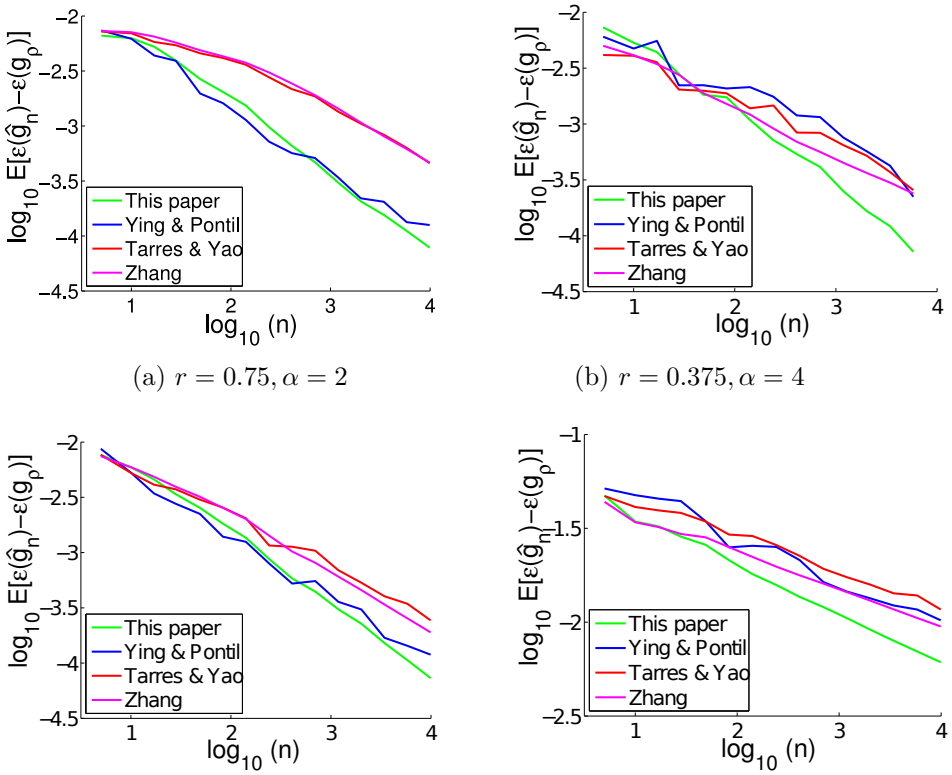


Figure 2: Optimal learning rate $\Gamma_{\text{best}}(n)$ for our algorithm in the finite horizon setting (plain magenta). The dashed green curve is a first order affine approximation of the second half of the magenta curve.

5.4. *Comparison to competing algorithms.* In this section, we compare the convergence rates of the four algorithms described in Section 4.5. We consider the different choices of (r, α) as described in Table 2 in order to go all over the different optimality situations. The main properties of each algorithm are described in Table 1. However we may note:

- For our algorithm, $\Gamma(n)$ is chosen accordingly with Corollary 2, with $\gamma_0 = \frac{1}{R^2}$.
- For Ying and Pontil's algorithm, accordingly to Theorem 6 in [8], we consider $\Gamma(n) = \gamma_0 n^{-\frac{2r}{2r+1}}$. We choose $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed $\frac{r}{64(1+R^4)(2r+1)}$.
- For Tarrès and Yaol's algorithm, we refer to Theorem C in [9], and consider $\Gamma(n) = a\,(n_0+n)^{-\frac{2r}{2r+1}}$ and $\Lambda(n) = \frac{1}{a}\,(n_0+n)^{-\frac{1}{2r+1}}$. The theorem is stated for all $a \geqslant 4$: we choose $a = 4$.
- For Zhangl's algorithm, we refer to Part 2.2 in [8], and choose $\Gamma(n) = \gamma_0 n^{-\frac{2r}{2r+1}}$ with $\gamma_0 = \frac{1}{R^2}$ which behaves better than the proposed choice $\frac{1}{4(1+R^2)}$.

Finally, we sum up the rates that were both predicted and derived for the four algorithms in the four cases for (α, δ) in Table 3. It appears that (a) we approximatively match the predicted rates in most cases (they would if n was larger), (b) our rates improve on existing work.



(a) $r = 0.75, \alpha = 2$ (b) $r = 0.375, \alpha = 4$

(c) $r = 1.25, \alpha = 2$ (d) $r = 0.125, \alpha = 4$

Figure 3: Comparison between algorithms. We have chosen parameters in each algorithm accordingly with description in Section 4.5, especially for the choices of γ_0. The y-axis is $\log_{10}\left(\mathbb{E}[\varepsilon(\hat{g}_n) - \varepsilon(g_\rho)]\right)$, where the final output $\hat{g}_n$ may be either $\overline{g}_n$ (This paper, Zhang) or g_n(Ying & Pontil, Yao & Tarres). This expectation is computed by averaging over 15 independent samples.

	$r = 0.75$ $\alpha = 2$	$r = 0.375$ $\alpha = 4$	$r = 1.25$ $\alpha = 2$	$r = 0.125$ $\alpha = 4$
Predicted rate (our algo.) Effective rate (our algo.)	**-0.75** **-0.7**	**-0.75** **-0.71**	**-0.8** **-0.69**	**-0.375** **-0.29**
Predicted rate (YP) Effective rate (YP)	-0.6 -0.53	-0.43 -0.5	-0.71 -0.63	-0.2 -0.22
Predicted rate (TY) Effective rate (TY)	-0.6 -0.48	 -0.39	 -0.43	 -0.2
Predicted rate (Z) Effective rate (Z)	 -0.53	-0.43 -0.43	 -0.41	-0.2 -0.21

TABLE 3

Predicted and effective rates (asymptotic slope of the log-log *plot) for the four different situations. We leave empty cases when the set-up does not come with existing guarantees: most algorithms seem to exhibit the expected behaviour even in such cases.*

6. Conclusion. In this paper, we have provided an analysis of averaged unregularized stochastic gradient methods for kernel-based least-squares regression. Our novel analysis allowed us to consider larger step-sizes, which in turn lead to optimal estimation rates for many settings of eigenvalue decay of the covariance operators and smoothness of the optimal prediction function. Moreover, we have worked on a more general setting than previous work, that includes most interesting cases of positive definite kernels.

Our work can be extended in a number of interesting ways: First, (a) we have considered results in expectation; following the higher-order moment bounds from [13] in the Euclidean case, we could consider higher-order moments, which in turn could lead to high-probability results or almost-sure convergence. Moreover, (b) while we obtain optimal convergence rates for a particular regime of kernels/objective functions, using different types of averaging (i.e., non uniform) may lead to optimal rates in other regimes. Besides, (c) following [13], we could extend our results for infinite-dimensional least-squares regression to other smooth loss functions, such as for logistic regression, where an online Newton algorithm with the same running-time complexity would also lead to optimal convergence rates. Finally, (d) the running-time complexity of our stochastic approximation procedures is still quadratic in the number of samples n, which is unsatisfactory when n is large; by considering reduced set-methods [31, 32, 6], we hope to able to obtain a complexity of $O(d_n n)$, where d_n is such that the convergence rate is $O(d_n/n)$, which would extend the Euclidean space result, where d_n is constant equal to the dimension.

Acknowledgements. This work was partially supported by the European Research Council (SIERRA Project). We thank Nicolas Flammarion for helpful discussions.

Appendices

In Appendix A, we provide proofs of the propositions from Section 2 that provide the Hilbert space set-up for kernel-based learning, while in Appendix B, we prove convergence rates for the least-mean-squares algorithm.

A. Reproducing kernel Hilbert spaces. In this appendix, we provide proofs of the results from Section 2 that provide the RHKS space set-up for kernel-based learning. See [20, 4, 34] for further properties of RKHSs.

We consider a reproducing kernel Hilbert space $\mathcal{H}$ with kernel K on space $\mathcal{X}$ as defined in Section 2.1. Unless explicitly mentioned, we do not make any topological assumption on $\mathcal{X}$.

As detailed in Section 2.2 we consider a set $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ and a distribution ρ on $\mathcal{X} \times \mathcal{Y}$. We denote ρ_X the marginal law on the space $\mathcal{X}$. In the following, we use the notation (X, Y) for a random variable following the law ρ. We define spaces $L^2_{\rho_X}, \mathcal{L}^2_{\rho_X}$ and the canonical projection p. In the following we further assume that $L^2_{\rho_X}$ is separable, an assumption satisfied in most cases.

We remind our assumptions:

(A1) $\mathcal{H}$ is a separable RKHS associated with kernel K on a space $\mathcal{X}$.
(A2) $\mathbb{E}\left[K(X, X)\right]$ and $\mathbb{E}[Y^2]$ are finite.

Assumption **(A2)** ensures that every function in $\mathcal{H}$ is square-integrable, that is, if $\mathbb{E}[K(X, X)] < \infty$, then $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$. Indeed, we have:

PROPOSITION 10. *Assume **(A1)**.*

1. *If $\mathbb{E}[K(X, X)] < \infty$, then $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$.*
2. *If $\sup_{x \in \mathcal{X}} K(x, x) < \infty$, then any function in $\mathcal{H}$ is bounded.*

PROOF. Under such condition, by Cauchy-Schwartz inequality, any function $f \in \mathcal{H}$ is either bounded or integrable:

$$\begin{aligned} |f(x)|^2 &\leqslant \|f\|_K^2 K(x, x) \leqslant \|f\|_K^2 \sup_{x \in \mathcal{X}} K(x, x), \\ \int_{\mathcal{X}} |f(x)|^2 d\rho_X(x) &\leqslant \|f\|_K^2 \int_{\mathcal{X}} K(x, x) d\rho_x(x). \end{aligned}$$

□

The assumption $\mathbb{E}[K(X, X)] < \infty$ seems to be the weakest assumption to make, in order to have at least $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$. However they may exist functions $f \in \mathcal{H} \setminus \{0\}$ such that $\|f\|_{\mathcal{L}^2_{\rho_X}} = 0$. However under stronger assumptions (see Section A.5) we may identify $\mathcal{H}$ and $p(\mathcal{H})$.

A.1. *Properties of the minimization problem.* We are interested in minimizing the following quantity, which is the *prediction error* of a function f, which may be rewritten as follows with dot-products in $L^2_{\rho_X}$:

$$\begin{aligned}
\varepsilon(f) &= \frac{1}{2}\mathbb{E}\left[(f(X)-Y)^2\right] \\
&= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \int_{\mathcal{X}\times\mathcal{Y}} f(x) y d\rho(x,y) + c \\
&= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \int_{\mathcal{X}} f(x)\left(\int_{\mathcal{Y}} y d\rho_{Y|X=x}(y)\right) d\rho_{|X}(x) + c \\
&= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \left\langle f, \int_{\mathcal{Y}} y d\rho_{Y|X=\cdot}(y)\right\rangle_{L^2_{\rho_X}} + c \qquad \text{(A.1)} \\
&= \frac{1}{2}\|f\|^2_{L^2_{\rho_X}} - \langle f, \mathbb{E}\left[Y|X=\cdot\right]\rangle_{L^2_{\rho_X}} + c
\end{aligned}$$

Notice that the problem may be re-written, if f is in $\mathcal{H}$, with dot-products in $\mathcal{H}$:

$$\begin{aligned}
\varepsilon(f) &= \frac{1}{2}\mathbb{E}[f(X)^2] - \langle f, \mathbb{E}[Y K_X]\rangle_K + \frac{1}{2}\mathbb{E}[Y^2] \\
&= \frac{1}{2}\langle f, \Sigma f\rangle_K - \langle f, \mu\rangle_K + c.
\end{aligned}$$

Interpretation: Under the form (A.1), it appears to be a minimisation problem in a Hilbert space of the sum of a continuous coercive function and a linear one. Using Lax-Milgramm and Stampachia theorems [17] we can conclude with the following proposition, which implies Prop. 1 in Section 2:

PROPOSITION 11 ($g_\rho, g_{\mathcal{H}}$). *Assume **(A1-2)**. We have the following points:*

1. *There exists a unique minimizer over the space $L^2_{\rho_X}$. This minimizer is the regression function $g_\rho : x \mapsto \int_{\mathcal{Y}} y d\rho_{Y|X=x}(y)$ (Lax-Milgramm).*
2. *For any non empty closed convex set, there exists a unique minimizer (Stampachia). As a consequence, there exists a unique minimizer:*

$$g_{\mathcal{H}} = \arg\min_{f\in\overline{p(\mathcal{H})}} \mathbb{E}\left[(f(X)-Y)^2\right]$$

over $\overline{p(\mathcal{H})}$. $g_{\mathcal{H}}$ is the orthogonal projection over g_ρ over $\overline{p(\mathcal{H})}$, thus satisfies the following equality: for any $\varepsilon \in \overline{H}$:

$$\mathbb{E}\left[(g_{\mathcal{H}}(X)-Y)\varepsilon(X)\right] = 0 \qquad \text{(A.2)}$$

A.2. *Covariance Operator.* We defined operators $\Sigma, \mathcal{T}, T$ in Section 2.4. We here state the main properties of these operators, then prove the two main decompositions stated in Propositions 2 and 3.

PROPOSITION 12 (Properties of Σ). *Assume **(A1-2)**.*

1. *Σ is well defined (that is for any $f \in \mathcal{H}$, $z \mapsto \mathbb{E}f(X)K(X,z)$ is in $\mathcal{H}$).*
2. *Σ is a continuous operator.*
3. *$\mathrm{Ker}(\Sigma) = \{f \in \mathcal{H} \text{ s.t. } \|f\|_{L^2_{\rho_X}} = 0\}$. Actually for any $f \in \mathcal{H}$, $\langle f, \Sigma f\rangle_K = \|f\|_{L^2_{\rho_X}}$.*
4. *Σ is a self-adjoint operator.*

PROOF. 1. for any $x \in \mathcal{X}$, $f(x)K_x$ is in $\mathcal{H}$. To show that the integral $\int_{x\in\mathcal{X}} f(x)K_x$ is converging, it is sufficient to show the is is absolutely converging in $\mathcal{H}$, as absolute convergence implies convergence in any Banach space[7] (thus any Hilbert space). Moreover:

$$\begin{aligned}
\int_{x\in\mathcal{X}} \|f(x)K_x\|_K &\leqslant \int_{x\in\mathcal{X}} |f(x)|\langle K_x, K_x\rangle_K^{1/2} \\
&\leqslant \int_{x\in\mathcal{X}} |f(x)|K(x,x)^{1/2} d\rho_X(x) \\
&\leqslant \left(\int_{x\in\mathcal{X}} f(x)^2 d\rho_X(x)\right)^{1/2} \left(\int_{x\in\mathcal{X}} K(x,x) d\rho_X(x)\right)^{1/2} \\
&< \infty,
\end{aligned}$$

under assumption $\mathbb{E}[K(X,X)] < \infty$ (**(A2)**).

2. For any $f \in \mathcal{H}$, we have

$$\begin{aligned}
\|\Sigma f\|_K &= \langle \Sigma f, \Sigma f\rangle_K = \int_{x\in\mathcal{X}} (\Sigma f)(x) f(x) d\rho_X(x) \\
&= \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{X}} f(y)K(x,y)d\rho_X(y)\right) f(x) d\rho_X(x) \\
&= \int_{x,y\in\mathcal{X}^2} \langle f, K_x\rangle_K \langle f, K_y\rangle_K \langle K_y, K_x\rangle_K d\rho_X(x) d\rho_X(y) \\
&\leqslant \int_{x,y\in\mathcal{X}^2} \|f\|_K \|K_x\|_K \|f\|_K \|K_y\|_K \|K_x\|_K \|K_y\|_K d\rho_X(x) d\rho_X(y) \\
&\qquad\qquad \text{by Cauchy Schwartz,} \\
&\leqslant \|f\|_K^2 \left(\int_{x\in\mathcal{X}^2} \|K_x\|_K^2 d\rho_X(x)\right)^2 \\
&\leqslant \|f\|_K^2 \left(\int_{x\in\mathcal{X}^2} K(x,x) d\rho_X(x)\right)^2,
\end{aligned}$$

[7] A Banch space is a linear normed space which is complete for the distance derived from the norm.

which proves the continuity under assumption **(A2)**.

3. $\Sigma f = 0 \Rightarrow \langle f, \Sigma f \rangle = 0 \Rightarrow \mathbb{E}[f^2(X)] = 0$. Reciprocally, if $\|f\|_{L^2_{\rho_X}} = 0$, it is clear that $\|\Sigma f\|_{L^2_{\rho_X}} = 0$, then $\|\Sigma f\|_K = \mathbb{E}\left[f(X)(\Sigma f)(X)\right] = 0$, thus $f \in \mathrm{Ker}(T)$.
4. It is clear that $\langle \Sigma f, g \rangle = \langle f, \Sigma g \rangle$.

□

PROPOSITION 13 (Properties of $\mathcal{T}$). *Assume **(A1-2)**. $\mathcal{T}$ satisfies the following properties:*

1. *$\mathcal{T}$ is a well defined, continuous operator.*
2. *For any $f \in \mathcal{H}$, $\mathcal{T}(\tilde{f}) = \Sigma f$, $\|\mathcal{T} f\|_K^2 = \int_{x,y \in \mathcal{X}^2} f(y) f(x) K(x,y) d\rho_X(y) d\rho_X(x)$.*
3. *The image of T is a subspace of $\mathcal{H}$.*

PROOF. It is clear that $\mathcal{T}$ is well defined, as for any class $\tilde{f}$, $\int_{\mathcal{X}} f(t)\, K_t\, d\rho_X(t)$ does not depend on the representer f, and is converging in $\mathcal{H}$ (which is the third point), just as in the previous proof. The second point results from the definitions. Finally for continuity, we have:

$$\begin{aligned}
\|\mathcal{T} f\|_K^2 &= \langle \mathcal{T} f, \mathcal{T} f \rangle_K \\
&= \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} f(y) f(x) K(x,y) d\rho_X(y) d\rho_X(x) \\
&\leqslant \left(\int_{x \in \mathcal{X}^2} |f(x) K(x,x)^{1/2}| d\rho_X(x) \right)^2 \\
&\leqslant \left(\int_{x \in \mathcal{X}} f(x)^2 d\rho_X(x) \right) \left(\int_{x \in \mathcal{X}} K(x,x) d\rho_X(x) \right) \leqslant C \|f\|_{L^2_{\rho_X}}^2 .
\end{aligned}$$

□

We now state here a simple lemma that will be useful later:

LEMMA 1. *Assume **(A1)**.*

1. *$\mathbb{E}\left[k(X,X)\right] < \infty \Rightarrow \int_{x,y \in \mathcal{X}} k(x,y)^2 d\rho_X(x) d\rho_X(y) < \infty$.*
2. *$\mathbb{E}\left[|k(x,y)|\right] < \infty \Rightarrow \int_{x,y \in \mathcal{X}} k(x,y)^2 d\rho_X(x) d\rho_X(y) < \infty$.*

PROPOSITION 14 (Propeties of T). *Assume **(A1-2)**. T satisfies the following properties:*

1. *T is a well defined, continuous operator.*
2. *The image of T is a subspace of $p(\mathcal{H})$.*
3. *T is a self-adjoint semi definite positive operator in the Hilbert space $L^2_{\rho_X}$.*

PROOF. $T = p \circ \mathcal{T}$ is clearly well defined, using the arguments given above. Moreover:

$$\begin{aligned}
\|Tf\|^2_{L^2_{\rho_X}} &= \int_{x\in\mathcal{X}} \left(\int_{t\in X} K(x,t) f(t) d\rho_X(t) \right)^2 d\rho_X(x) \\
&\leqslant \left(\int_{x\in\mathcal{X}} \int_{t\in X} K(x,t)^2 d\rho_X(t) d\rho_X(x) \right) \left(\int_{t\in\mathcal{X}} f^2(t) d\rho_X(t) \right) \text{ by C.S.} \\
&\leqslant C\|f\|_{\mathcal{L}^2_{\rho_X}} \quad \text{by Lemma 1,}
\end{aligned}$$

which is continuity[8]. Then by Proposition 13, $\mathrm{Im}(Td) \subset p(\mathrm{Im}(\mathcal{T})) \subset p(\mathcal{H})$. Finally, for any $f, g \in \mathcal{L}^2_{\rho_X}$,

$$\begin{aligned}
\langle f, Tg\rangle_{\mathcal{L}^2_{\rho_X}} &= \int_{\mathcal{X}} f(x)\ Tg(x) d\rho_X(x) \\
&= \int_{\mathcal{X}} f(x) \left(\int_{\mathcal{X}} g(t) K(x,t) d\rho_X(t) \right) d\rho_X(x) \\
&= \int_{\mathcal{X}\times\mathcal{X}} f(x) g(t) K(x,t) d\rho_X(t) d\rho_X(x) = \langle Tf, g\rangle_{\mathcal{L}^2_{\rho_X}}.
\end{aligned}$$

and $\langle f, Tf\rangle_{\mathcal{L}^2_{\rho_X}} \geq 0$ as a generalisation of the positive definite property of K. □

In order to show the existence of an eigenbasis for T, we now show that is trace-class.

PROPOSITION 15 (Compactness of the operator). *We have the following properties:*

1. *Under **(A2)**, T is a trace class operator[9]. As a consequence, it is also a Hilbert-Schmidt operator[10].*
2. *If $K \in L^2(\rho_X \times \rho_X)$ then T is a Hilbert-Schmidt operator.*
3. *Any Hilbert-Schmidt operator is a compact operator.*

PROOF. Proofs of such facts may be found in [17, 41]. Formally, with

[8] We could also use the continuity of $p : \mathcal{H} \to L^2_{\rho_X}$.

[9] Mimicking the definition for matrices, a bounded linear operator A over a separable Hilbert space H is said to be in the trace class if for some (and hence all) orthonormal bases $(e_k)_k$ of H the sum of positive terms $\mathrm{tr}|A| := \sum_k \langle (A^*A)^{1/2} e_k, e_k\rangle$ is finite.

[10] A Hilbert-Schmidt operator is a bounded operator A on a Hilbert space H with finite Hilbert–Schmidt norm: $\|A\|^2_{\mathrm{HS}} = \mathrm{tr}|(A^*A)| := \sum_{i\in I} \|Ae_i\|^2$.

$(\phi_i)_i$ an Hilbertian basis in $L^2_{\rho_X}$:

$$\begin{aligned}
\mathbb{E}\left[K(X,X)\right] &= \mathbb{E}\left[\langle K_x, K_x\rangle_K\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty}\langle K_x,\phi_i\rangle_K^2\right] \quad \text{by Parseval equality,} \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\langle K_x,\phi_i\rangle_K^2\right] \\
&= \sum_{i=1}^{\infty}\langle T\phi_i,\phi_i\rangle_K = \mathrm{tr}(T).
\end{aligned}$$

□

COROLLARY 6. *We have thus proved that under **(A1)** and **(A2)**, the operator T may be reduced in some Hilbertian eigenbasis: the fact that T is self-adjoint and compact implies the existence of an orthonormal eigensystem (which is an Hilbertian basis of $L^2_{\rho_X}$).*

This is a consequence of a very classical result, see for example [17].

DEFINITION 4. *The null space $Ker(T) := \left\{f \in L^2_{\rho_X} \text{ s.t. } Tf = 0\right\}$ may not be $\{0\}$. We denote by S an orthogonal supplementary of $Ker(T)$.*

Proposition 2 is directly derived from a slightly more complete Proposition 16 below:

PROPOSITION 16 (Eigen-decomposition of T). *Under **(A1)** and **(A2)**, T is a bounded self adjoint semi-definite positive operator on $L^2_{\rho_X}$, which is trace-class. There exists[11] a Hilbertian eigenbasis $(\phi_i)_{i\in I}$ of the orthogonal supplement S of the null space $\mathrm{Ker}(T)$, with summable eigenvalues $(\mu_i)_{i\in I}$. That is:*

- *$\forall i \in I$, $T\phi_i = \mu_i\phi_i$, $(\mu_i)_i$ strictly positive non increasing (or finite) sequence such that $\sum_{i\in I}\mu_i < \infty$.*
- *$L^2_{\rho_X} = \mathrm{Ker}(T) \overset{\perp}{\oplus} S$.*

We have[12]: $S = \overline{span\{\phi_i\}} = \left\{\sum_{i=1}^{\infty} a_i\phi_i \text{ s.t. } \sum_{i=1}^{\infty} a_i^2 < \infty\right\}$. Moreover:

$$S = \overline{p(\mathcal{H})}. \tag{A.3}$$

[11] S is stable by T and $T : S \to S$ is a self adjoint compact positive operator.

[12] We denote by $\mathrm{span}(A)$ the smallest linear space which contains A, which is in such a case the set of all finite linear combinations of $(\phi_i)_{i\in I}$.

PROOF. For any $i \in I$, $\phi_i = \frac{1}{\mu_i} L_K \phi_i \in p(\mathcal{H})$. Thus $\text{span}\{\phi_i\} \subset p(\mathcal{H})$, thus $S = \overline{\text{span}\{\phi_i\}} \subset \overline{p(\mathcal{H})}$. Moreover, using the following Lemma, $p(\mathcal{H}) \subset \text{Ker}(T)^\perp = S$, which concludes the proof, by taking the closures. □

LEMMA 2. *We have the following points:*

- *if $T^{1/2} f = 0$ in $L^2_{\rho_X}$, then $Tf = 0$ in $\mathcal{H}$.*
- *$p(\mathcal{H}) \subset \text{Ker}(T)^\perp$.*

PROOF. We first notice that if $T^{1/2} f = 0$ in $L^2_{\rho_X}$, then $\mathcal{T} f = 0$ in $\mathcal{H}$: indeed[13]

$$\begin{aligned} \|Tf\|^2_{\mathcal{H}} &= \left\langle \int_{\mathcal{X}} f(x) K_x d\rho_X(x), \int_{\mathcal{X}} f(y) K_y d\rho_X(y) \right\rangle_K \\ &= \int_{\mathcal{X}^2} f(x) f(y) K(x,y) d\rho_X(x) d\rho_X(y) \\ &= \langle f, Tf \rangle_{L^2_{\rho_X}} = 0 \text{ if } Tf = 0 \text{ in } L^2_{\rho_X}. \end{aligned}$$

Moreover $\mathcal{H}$ is the completed space of $\text{span}\{K_x, x \in \mathcal{X}\}$, with respect to $\|\cdot\|_K$ and for all $x \in \mathcal{X}$, for all $\psi_k \in \text{Ker}(T)$:

$$\begin{aligned} \langle p(K_x), \psi_k \rangle_{L^2_{\rho_X}} &= \int_{\mathcal{X}} K_x(y) \psi_K(y) d\rho_X(y) = (T\psi_k)(x), \\ \text{however,} \quad T\psi_k &=_{L^2_{\rho_X}} 0 \quad \Rightarrow T\psi_k =_{\mathcal{H}} 0 \quad \forall x \in \mathcal{X} \Rightarrow T\psi_k(x) = 0. \end{aligned}$$

As a consequence, $\text{span}\{p(K_x), x \in \mathcal{X}\} \subset \text{Ker}(T)^\perp$. We just have to show that $\overline{\text{span}\{p(K_x), x \in \mathcal{X}\}} = p(\mathcal{H})$, as $\text{Ker}(T)^\perp$ is a closed space. It is true as for any $\tilde{f} \in p(H), f \in \mathcal{H}$ there exists $f_n \subset \text{span}\{K_x, x \in \mathcal{X}\}$ such that $f_n \overset{\mathcal{H}}{\to} f$, thus $p(f_n) \to \tilde{f}$ in $L^2_{\rho_X}$[14]. Finally we have proved that $p(\mathcal{H}) \subset \text{Ker}(T)^\perp$. □

Similarly, Proposition 3 is derived from Proposition 17 below:

[13] In other words, we the operator defined below $T^{1/2}$

$$\begin{aligned} T^{1/2} f &=_{L^2_{\rho_X}} 0 \\ \mathcal{T} f &=_{\mathcal{H}} \Sigma^{1/2}(\mathcal{T}^{1/2} f) \\ \|\mathcal{T} f\|^2_K &= \|\Sigma^{1/2}(\mathcal{T}^{1/2} f)\|^2_K = \|(\mathcal{T}^{1/2} f)\|^2_{L^2_{\rho_X}} = 0 \\ {}^{H}Tf &=_{\mathcal{H}} 0. \end{aligned}$$

[14] $\|f_n - f\|_{L^2_{\rho_X}} = \|\Sigma^{1/2}(f_n - f)\|_K \to 0$ as Σ continuous.

PROPOSITION 17 (Decomposition of Σ). *Under **(A1)** and **(A2)**, $\mathrm{Im}(\mathcal{T}) \subset \mathcal{H}$, that is, for any $f \in L^2_{\rho_X}$, $\mathcal{T}f \in \mathcal{H}$. Moreover, for any $i \in I$, $\phi_i^H = \frac{1}{\mu_i}\mathcal{T}\phi_i \in H$ is a representant for the equivalence class ϕ_i. Moreover $\left(\mu_i^{1/2}\phi_i^H\right)_{i\in I}$ is an orthonormal eigein-system of $\mathscr{S}$ That is:*

- $\forall i \in I,\ \Sigma\phi_i^H = \mu_i\phi_i^H$.
- $\left(\mu_i^{1/2}\phi_i^H\right)_{i\in I}$ *is an orthonormal family in* $\mathscr{S}$.

We thus have:

$$\mathscr{S} = \left\{\sum_{i\in I} a_i\phi_i^H \ \ s.t. \ \sum_{i\in I}\frac{a_i^2}{\mu_i} < \infty\right\}.$$

Moreover $\mathscr{S}$ is the orthogonal supplement of the null space Ker(Σ):

$$\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}.$$

PROOF. The family $\phi_i^H = \frac{1}{\mu_i}T\phi_i$ satisfies:

- $\widetilde{\phi_i^H} = \phi_i$ (in $L^2_{\rho_X}$),
- $\phi_i^H \in \mathscr{S}$,
- $T\phi_i^H = \mu_i\phi_i$ in $L^2_{\rho_X}$,
- $\mathcal{T}\phi_i^H = \Sigma\phi_i^H = \mu_i\phi_i^H$ in $\mathcal{H}$.

All the points are clear: indeed for example $\Sigma\phi_i^H = T\phi_i = \mu_i\phi_i^H$. Moreover, we have that:

$$\begin{aligned}\|\phi_i\|^2_{L^2_{\rho_X}} = \|\phi_i^H\|^2_{L^2_{\rho_X}} &= \langle\phi_i^H, \Sigma\phi_i\rangle_K \text{ by Proposition 3}\\ &= \mu_i\|\phi_i^H\|_K^2\\ &= \|\sqrt{\mu_i}\phi_i^H\|_K^2\end{aligned}$$

That means that $(\sqrt{\mu_i}\phi_i^H)_i$ is an orthonormal family in $\mathcal{H}$.

Moreover, $\mathscr{S}$ is defined as the completion for $\|\cdot\|_K$ of this orthonormal family, which gives $\mathscr{S} = \left\{\sum_{i\in I} a_i\phi_i^H \text{ s.t. } \sum_{i\in I}\frac{a_i^2}{\mu_i} < \infty\right\}$.

To show that $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$, we use the following sequence of arguments:

- First, as Σ is a continuous operator, $\mathrm{Ker}(\Sigma)$ is a closed space in $\mathcal{H}$, thus $\mathcal{H} = \mathrm{Ker}(\Sigma) \overset{\perp}{\oplus} (\mathrm{Ker}(\Sigma))^\perp$.
- $\mathrm{Ker}(\Sigma) \subset (\mathcal{T}^{1/2}(S))^\perp$: indeed for all $f \in \mathrm{Ker}(\Sigma)$, $\langle f, \phi_i^{\mathcal{H}}\rangle = \frac{1}{\mu_i}\langle f, \Sigma\phi_i^{\mathcal{H}}\rangle = \frac{1}{\mu_i}\Sigma\langle f, \phi_i^{\mathcal{H}}\rangle = 0$, and as a consequence for any $f \in \mathrm{Ker}(\Sigma), g \in \mathcal{T}^{1/2}(S)$,

there exists $(g_n) \subset \text{span}(\phi_i^H)$ s.t. $g_n \overset{\mathcal{H}}{\to} g$, thus $0 = \langle g_n, f\rangle_{\mathcal{H}} \to \langle f, g\rangle$ and finally $f \in (\mathcal{T}^{1/2}(S))^\perp$. Equivalently $\mathcal{T}^{1/2}(S) \subset (\text{Ker}(\Sigma))^\perp$.

- $(\mathcal{T}^{1/2}(S))^\perp \subset \text{Ker}(\Sigma)$. For any i, $\phi_i^H \in \mathcal{T}^{1/2}(S)$. If $f \in (\mathcal{T}^{1/2}(S))^\perp$, then $\langle p(f), \phi_i\rangle_{L^2_{\rho_X}} = \langle f, \mathcal{T}\phi_i\rangle_{\mathcal{H}} = 0$. As a consequence $p(f) \in p(\mathcal{H}) \cap \text{Ker}(T) = \{0\}$, thus $f \in \text{Ker}(\Sigma)$. That is $(\mathcal{T}^{1/2}(S))^\perp \subset \text{Ker}(\Sigma)$. Equivalently $\text{Ker}(\Sigma)^\perp \subset (\mathcal{T}^{1/2}(S))$.
- Combining these points: $\mathcal{H} = \text{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$.

□

We have two decompositions of $\mathcal{L}^2_{\rho_X} = \text{Ker}(T) \overset{\perp}{\oplus} S$ and $\mathcal{H} = \text{Ker}(\Sigma) \overset{\perp}{\oplus} \mathscr{S}$. They happen to be related through the mapping $\mathcal{T}^{1/2}$, which we now define.

A.3. *Properties of T^r, $r > 0$.* We defined operators T^r, $r > 0$ and $\mathcal{T}^r$, $r \geqslant 1/2$ in Section 2.4 in Definitions 2,3.

PROPOSITION 18 (Properties of T^r, $\mathcal{T}^r$).

- *T^r is well defined for any $r > 0$.*
- *$\mathcal{T}^r$ is well defined for any $r \geqslant \frac{1}{2}$.*
- *$\mathcal{T}^{1/2} : S \to \mathscr{S}$ is an isometry.*
- *Moreover $\text{Im}(T^{1/2}) = p(\mathcal{H})$. That means $T^{1/2} : S \to p(\mathcal{H})$ is an isomorphism.*

PROOF. $\underline{T^r \text{ is well defined for any } r > 0.}$
$S = \{\sum_{i=1}^\infty a_i\phi_i \text{ s.t. } \sum_{i=1}^\infty a_i^2 < \infty\}$. For any sequence $(a_i)_{i\in I}$ such that $\sum_{i=1}^\infty a_i^2 < \infty$, $T^r(\sum a_i\phi_i) = \sum_i \mu_i^r a_i\phi_i$ is a converging sum in the Hilbert space $L^2_{\rho_X}$ (as $(\mu_i)_{i\in I}$ is bounded thus $\sum_i \mu_i^r a_i\phi_i$ satisfies Cauchy is criterion: $\|\sum_{i=n}^p \mu_i^r a_i\phi_i\|^2 \leqslant \mu_0^r \left(\sum_{i=n}^p a_i^2\right)^{1/2}$). And Cauchy is criterion implies convergence in Hilbert spaces.

$\underline{\mathcal{T}^r \text{ is well defined for any } r \geqslant \frac{1}{2}.}$
We have shown that $(\sqrt{\mu_i}\phi_i^H)_i$ is an orthonormal family in $\mathcal{H}$. As a consequence (using the fact that (μ_i) is a bounded sequence), for any sequence $(a_i)_i$ such that $\sum a_i^2 < \infty$, $\sum_i \mu_i^r a_i\phi_i^H$ satisfies Cauchy is criterion thus is converging in $\mathcal{H}$ as $\|\sum_{i\in I'} \mu_i^r a_i\phi_i^H\|_K = \sum_{i\in I'} \mu_i^{r-1/2} a_i^2 \leqslant \mu_0^{r-1/2} \sum_{i\in I'} a_i^2 < \infty$. (We need $r \geqslant 1/2$ of course).

$\underline{\mathcal{T}^{1/2} : S \to \mathscr{S} \text{ is an isometry.}}$
Definition has been proved. Surjectivity in $\mathscr{S}$ is by definition, as $\mathcal{T}^{1/2}(S) = \left\{\sum_{i\in I} a_i\phi_i^H \text{ s.t. } \sum_{i\in I} \frac{a_i^2}{\mu_i} < \infty\right\}$. Moreover, the operator is clearly injective

as for any $f \in S$, $Tf \neq 0$ in $L^2_{\rho_X}$ thus $Tf \neq 0$ in $\mathcal{H}$. Moreover for any $f = \sum_{i=1}^{\infty} a_i \phi_i \in S$, $\|Tf\|_K^2 = \|\sum_{i=1}^{\infty} a_i \sqrt{\mu_i}\phi_i\|_K^2 = \sum_{i=1}^{\infty} a_i^2 = \|f\|^2_{\mathcal{L}^2_{\rho_X}}$, which is the isometrical property.

It must be noticed that we cannot prove surjectivity in $\mathcal{H}$[15], that is without our "strong assumptions". However we will show that operator $T^{1/2}$ is surjective in $p(\mathcal{H})$.

$\underline{\mathrm{Im}(T^{1/2}) = p(\mathcal{H}). \text{ That means } T^{1/2} : S \to p(\mathcal{H}) \text{ is an isomorphism.}}$
$\mathrm{Im}(T^{1/2}) = p(\mathrm{Im}(\mathcal{T}^{1/2})) = p(\mathscr{S})$. Moreover $p(\mathcal{H}) = p(\mathrm{Ker}(\Sigma) \oplus \mathscr{S}) = p(\mathscr{S})$. Consequently $\mathrm{Im}(T^{1/2}) = p(\mathcal{H})$. Moreover $T^{1/2} : S \to L^2_{\rho_X}$ is also injective, which give the isomorphical character.

Note that it is clear that $T^{1/2}(S) \subset p(\mathcal{H})$ and that for any $x \in \mathcal{X}$, $p(K_x) \in T^{1/2}(S)$ indeed $p(K_x) = \sum_{i=1}^{\infty} \langle K_x, \phi_i \rangle_{L^2_{\rho_X}} \phi_i = \sum_{i=1}^{\infty} \mu_i \phi_i^H(x) \phi_i$, with $\sum_{i=1}^{\infty} \frac{\left(\mu_i \phi_i^H(x)\right)^2}{\mu_i} = \sum_{i=1}^{\infty} \mu_i \phi_i^H(x)^2 < \infty$, as $K(x,x) = \sum_{i=1}^{\infty} \mu_i \phi_i^H(x)^2$ □

Finally, it has appeared that S and $\mathscr{S}$ may be identified via the isometry $\mathcal{T}^{1/2}$. We conclude by a proposition which sums up the properties of the spaces $\mathcal{T}^r(L^2_{\rho_X})$.

PROPOSITION 19. *The spaces $T^r(L^2_{\rho_X}), r > 0$ satisfy:*

$$\begin{aligned}
\forall r \geqslant r' > 0, \quad T^r\left(L^2_{\rho_X}\right) &\subset T^{r'}\left(L^2_{\rho_X}\right) \\
\forall r > 0, \quad \overline{T^r\left(L^2_{\rho_X}\right)} &= S \\
T^{1/2}\left(L^2_{\rho_X}\right) &= p(\mathcal{H}) \\
\forall r \geqslant \frac{1}{2}, \quad T^r\left(L^2_{\rho_X}\right) &\subset p(\mathcal{H})
\end{aligned}$$

A.4. *Kernel decomposition.* We prove here Proposition 6.

PROOF. Considering our decomposition of $\mathcal{H} = \mathscr{S} \overset{\perp}{\oplus} \ker(\Sigma)$, an the fact the $(\sqrt{\mu_i}\phi_i^{\mathcal{H}})$ is a Hilbertian eigenbasis of $\mathscr{S}$, we have for any $x \in \mathcal{X}$,

$$\begin{aligned}
K_x &= \sum_{i=1}^{\infty} \langle \sqrt{\mu_i}\phi_i^{\mathcal{H}}, K_x \rangle_{\mathcal{H}} \sqrt{\mu_i}\phi_i^{\mathcal{H}} + g_x \\
&= \sum_{i=1}^{\infty} \mu_i \phi_i^{\mathcal{H}}(x) \phi_i^{\mathcal{H}} + g_x
\end{aligned}$$

[15] It is actually easy to build a counter example, f.e. with a measure of "small" support (let is say $[-1,1]$), a Hilbert space of functions on $\mathcal{X} = [-5;5]$, and a kernel like $\min(0, 1 - |x-y|)$: $\mathrm{Im}\left(\mathcal{T}^{1/2}\right) \subset \{f \in \mathcal{H} \text{ s. t. } \mathrm{supp}(f) \subset [-2;2]\} \subsetneq \mathcal{H}$.

And as it has been noticed above this sum is converging in $\mathscr{S}$ (as in $\mathcal{H}$) because $\sum_{i=1}^{\infty} \frac{(\mu_i \phi_i^{\mathcal{H}}(x))^2}{\mu_i} = \sum_{i=1}^{\infty} \mu_i (\phi_i^{\mathcal{H}}(x))^2 = K(x,x) < \infty$. However, the convergence may not be absolute in $\mathcal{H}$. Our function g_x is in $\mathrm{Ker}(\Sigma)$, which means $\int_{y \in \mathcal{X}} g_x(y)^2 d\rho_X(y) = 0$.

And as a consequence, we have for all $x, y \in \mathcal{X}$,

$$K(x,y) = \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y) + g(x,y),$$

With $g(x,y) := g_x(y)$. Changing roles of x, y, it appears that $g(x,y) = g(y,x)$. And we have for all $x \in \mathcal{X}$, $\int_{\mathcal{X}} g(x,y)^2 d\rho_X(y) = 0$. Moreover, the convergence of the series is absolute

We now prove the following points

(a) $(\mathscr{S}, \|\cdot\|_{\mathcal{H}})$ is also an RKHS, with kernel $K^{\mathscr{S}} : (x,y) \mapsto \sum_{i \in I} \mu_i \phi_i^H(x) \phi_i^H(y)$
(b) given the decomposition above, almost surely the optimization problem in $\mathcal{S}$ and $\mathcal{H}$ have equivalent solutions.

(a) $(\mathscr{S}, \|\cdot\|_{\mathcal{H}})$ is a Hilbert space as a closed subspace of a Hilbert space. Then for any $x \in \mathcal{X}$: $K_x^{\mathscr{S}} := (y \mapsto K^{\mathscr{S}}(x,y)) = \sum_{i=1}^{\infty} \mu_i \phi_i^{\mathcal{H}}(x) \phi_i^{\mathcal{H}} \in \mathscr{S}$. Finally, for any $f \in \mathscr{S}$

$$\langle f, K_x^{\mathscr{S}} \rangle_{\mathcal{H}} = \langle f, K_x^{\mathscr{S}} + g_x \rangle_{\mathcal{H}} = \langle f, K_x \rangle_{\mathcal{H}} = f(x),$$

because $g_x \in \mathrm{Ker}(\Sigma) = \mathscr{S}^{\perp} \ni f$. Thus stands the reproducing property.

(b) We have that $p(\mathscr{S}) = p(\mathcal{H})$ and our best approximating function is a minimizer over this set. Moreover if $K_x^{\mathscr{S}}$ was used instead of K_x in our algorithm, both estimators are almost surely almost surely equal (i.e., almost surely in the same equivalence class). Indeed, at any step n, if we denote $g_n^{\mathscr{S}}$ the sequence built in $\mathscr{S}$ with $K^{\mathscr{S}}$, if we have $g_n^{\mathscr{S}} \overset{a.s.}{=} g_n$, then almost surely $g_n^{\mathscr{S}}(x_n) = g_n(x_n)$ and moreover $K_{x_n} \overset{a.s.}{=} K_{x_n}^S$. Thus almost surely, $g_{n+1} \overset{a.s.}{=} g_{n+1}^{\mathscr{S}}$.

□

A.5. *Alternative assumptions.* As it has been noticed in the paper, we have tried to minimize assumptions made on $\mathcal{X}$ and K. In this section, we review some of the consequences of such assumptions.

A.5.1. *Alternative assumptions.* The following have been considered previously:

1. Under the assumption that ρ *is a Borel probability measure* (with respect with some topology on $\mathbb{R}^d$) and $\mathcal{X}$ is a closed space, we may assume that $\mathrm{supp}(\rho) = \mathcal{X}$, where $\mathrm{supp}(\rho)$ is the smallest close space of measure one.

2. The assumption that K *is a Mercer kernel* ($\mathcal{X}$ compact, K continuous) has generally been made before [9, 16, 4, 8], but does not seem to be necessary here.
3. **(A2)** was replaced by the stronger assumption $\sup_{x\in\mathcal{X}} K(x,x) < \infty$ [9, 8, 7] and $|Y|$ bounded [9, 7] .

A.5.2. *Identification $\mathcal{H}$ and $p(\mathcal{H})$.* Working with mild assumptions has made it necessary to work with sub spaces of $L^2_{\rho_X}$, thus projecting $\mathcal{H}$ in $p(\mathcal{H})$. With stronger assumptions given above, the space $\mathcal{H}$ may be identified with $p(\mathcal{H})$.

Our problems are linked with the fact that a function f in $\mathcal{H}$ may satisfy both $\|f\|_{\mathcal{H}} \neq 0$ and $\|f\|_{L^2_{\rho_X}} = 0$.

- the "support" of ρ may not be $\mathcal{X}$.
- even if the support is $\mathcal{X}$, a function may be ρ-a.s. 0 but not null in $\mathcal{H}$.

Both these "problems" are solved considering the further assumptions above. We have the following Proposition:

PROPOSITION 20. *If we consider a Mercer kernel K (or even any continuous kernel), on a space $\mathcal{X}$ compact and a measure ρ_X on $\mathcal{X}$ such that* $\mathrm{supp}(\rho) = \mathcal{X}$ *then the map:*

$$\begin{array}{rcl} p : \mathcal{H} & \to & p(\mathcal{H}) \\ f & \mapsto & \tilde{f} \end{array}$$

is injective, thus bijective.

A.5.3. *Mercer kernel properties.* We review here some of the properties of Mercer kernels, especially Mercer's theorem which may be compared to Proposition 6.

PROPOSITION 21 (Mercer theorem). *Let $\mathcal{X}$ be a compact domain or a manifold, ρ a Borel measure on $\mathcal{X}$, and $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ a Mercer Kernel. Let λ_k be the k-th eigenvalue of T and Φ_k the corresponding eigenvectors. For all $x,t \in \mathcal{X}$, $K(x,t) = \sum_{k=1}^{\infty} \lambda_k \Phi_k(x)\Phi_k(t)$ where the convergence is absolute (for each $x,t \in \mathcal{X}^2$) and uniform on $\mathcal{X}\times\mathcal{X}$.*

The proof of this theorem is given in [42].

PROPOSITION 22 (Mercer Kernel properties). *In a Mercer kernel, we have that:*

1. $C_K := \sup_{x,t\in\mathcal{X}^2}(K(x,t)) < \infty$.
2. $\forall f \in \mathcal{H}$, f *is* C^0.
3. *The sum* $\sum \lambda_k$ *is convergent and* $\sum_{k=1}^{\infty} \lambda_k = \int_X K(x,x) \leqslant \rho(\mathcal{X})C_K$.
4. *The inclusion* $I_K : \mathcal{H} \to C(\mathcal{X})$ *is bounded with* $|||I_K||| \leqslant C_K^{1/2}$.
5. *The map*

$$\begin{aligned} \Phi : \mathcal{X} &\to \ell^2 \\ x &\mapsto (\sqrt{\lambda_k}\Phi_k(x))_{k\in\mathbb{N}} \end{aligned}$$

   is well defined, continuous, and satisfies $K(x,t) = \langle \Phi_k(x), \Phi_k(t)\rangle$.
6. *The space* $\mathcal{H}$ *is independent of the measure considered on* $\mathcal{X}$.

We can characterize $\mathcal{H}$ via the eigenvalues-eigenvectors:

$$\mathcal{H} = \left\{ f \in L^2_{\rho_X} | f = \sum_{k=1}^{\infty} a_k \Phi_k \text{ with } \sum_{k=1}^{\infty} \left(\frac{a_k}{\sqrt{\lambda_k}} \right)^2 < \infty \right\}.$$

Which is equivalent to saying that $T^{1/2}$ is an isomorphism between $L^2_{\rho_X}$ and $\mathcal{H}$. Where we have only considered $\lambda_k > 0$. It has no importance to consider the linear subspace S of $L^2_{\rho_X}$ spanned by the eigenvectors with non zero eigenvalues. However it changes the space $\overline{\mathcal{H}}$ which is in any case S, and is of some importance regarding the estimation problem.

B. Proofs. To get our results, we are going to derive from our recursion a new error decomposition and bound the different sources of error via algebraic calculations. We first make a few remarks on short notations that we will use in this part and difficulties that arise from the Hilbert space setting in Section B.1, then provide intuition via the analysis of a closely related recursion in Section B.2. We give in Sections B.3, B.4 the complete proof of our bound respectively in the finite horizon case (Theorem 2) and the online case (Theorem 3). We finally provide technical calculations of the main bias and variance terms in Section B.5.

B.1. *Preliminary remarks.* We remind that we consider a sequence of functions $(g_n)_{n\in\mathbb{N}}$ satisfying the system defined in Section 3.

$$\begin{aligned} g_0 &= 0 \text{ (the null function)}, \\ g_n &= \sum_{i=1}^{n} a_i K_{x_i}. \end{aligned}$$

With a sequence $(a_n)_{n\geqslant 1}$ such that for all n greater than 1 :

$$a_n := -\gamma_n(g_{n-1}(x_n) - y_n) = -\gamma_n \left(\sum_{i=1}^{n-1} a_i K(x_n, x_i) - y_i \right). \tag{B.1}$$

We output

$$\overline{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} \overline{g}_k. \tag{B.2}$$

We consider a representer $g_{\mathcal{H}} \in \mathcal{L}^2_{\rho_X}$ of $g_{\mathcal{H}}$ defined by Proposition 1. We accept to confuse notations as far as our calculations are made on $L^2_{\rho_X}$-norms, thus does not depend on our choice of the representer.

We aim to estimate :

$$\varepsilon(\overline{g}_n) - \varepsilon(g_{\mathcal{H}}) = \frac{1}{2}\|\overline{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}.$$

B.1.1. *Notations.* In order to simplify reading, we will use some shorter notations :

- For the covariance operator, we will only use Σ instead of $\Sigma, T, \mathcal{T}$,

Space :	$\mathcal{H}$
Observations :	$(x_n, y_n)_{n\in\mathbb{N}}$ i.i.d. $\sim \rho$
Best approximation function :	$g_{\mathcal{H}}$
Learning rate :	$(\gamma_i)_i$

All the functions may be split up the orthonormal eigenbasis of the operator $\mathcal{T}$. We can thus see any function as an infinite-dimensional vector, and operators as matrices. This is of course some (mild) abuse of notations if we are not in finite dimensions. For example, our operator Σ may be seen as $\mathrm{Diag}(\mu_i)_{1\leqslant i}$. Carrying on the analogy with the finite dimensional setting, a self adjoint operator, may be seen as a symmetric matrix.

We will have to deal with several "matrix products" (which are actually operator compositions). We denote :

$$\begin{aligned}
M(k,n,\gamma) &= \prod_{i=k}^{n}(I-\gamma K_{x_i}\otimes K_{x_i}) = (I-\gamma K_{x_k}\otimes K_{x_k})\cdots(I-\gamma K_{x_n}\otimes K_{x_n}) \\
M(k,n,(\gamma_i)_i) &= \prod_{i=k}^{n}(I-\gamma_i K_{x_i}\otimes K_{x_i}) \\
D(k,n,(\gamma_i)_i) &= \prod_{i=k}^{n}(I-\gamma_i \Sigma)
\end{aligned}$$

Remarks :

- As our operators may not commute, we use a somehow unusual convention by defining the products for any k, n, even with $k > n$, with $M(k,n,\gamma) = (I-\gamma K_{x_k}\otimes K_{x_k})(I-\gamma K_{x_{k-1}}\otimes K_{x_{k-1}})\cdots(I-\gamma K_{x_n}\otimes K_{x_n})$.

- We may denote $D(k,n,\gamma) = \prod_{i=k}^{n}(I-\gamma\Sigma)$ even if its clearly $(I-\gamma\Sigma)^{n-k+1}$ just in order to make the comparison between equations easier.

B.1.2. *On norms.* In the following, we will use constantly the following observation :

LEMMA 3. *Assume* ***A2-4*** *, let* $\eta_n = g_n - g_{\mathcal{H}}$, $\bar{\eta}_n = \overline{g}_n - g_{\mathcal{H}}$ *:*

$$\begin{aligned}\varepsilon(g_n) - \varepsilon(g_{\mathcal{H}}) &= \frac{1}{2}\langle \eta_n, \Sigma\eta_n\rangle = \frac{1}{2}\mathbb{E}\left[\langle x, g_n - g_{\mathcal{H}}\rangle^2\right]\left(:= \frac{1}{2}\|g_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right),\\ \varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}}) &= \frac{1}{2}\langle \bar{\eta}_n, \Sigma\bar{\eta}_n\rangle.\end{aligned}$$

B.1.3. *On symmetric matrices.* One has to be careful when using auto adjoint operators, especially when using the order $A \preccurlyeq B$ which means that $B - A$ is non-negative.

Some problems may arise when some self adjoint A, B do not commute, because then AB is not even in auto adjoint. It is also hopeless to compose such relations : for example $A \preccurlyeq B$ does not imply $A^2 \preccurlyeq B^2$ (while the opposite is true).

However, it is true that if $A \preccurlyeq B$, then for any C in $S_n(\mathbb{R})$, we have $C^t\,AC \preccurlyeq C^t\,BC$. We will often use this final point. Indeed for any x, $x^t\,(C^t\,BC - C^t\,AC)x = (Cx)^t(B-A)(Cx) \geqslant 0$.

B.1.4. *Notation.* In the proof, we may use, for any $x \in \mathcal{H}$:

$$\begin{aligned}\widetilde{K_x \otimes K_x} \,: L^2_{\rho_X} &\to \mathcal{H}\\ f &\mapsto f(x)\; K_x.\end{aligned}$$

We only consider functions $\mathcal{L}^2_{\rho_X}$, which are well defined at any point. The regression function is only almost surely defined but we will consider a version of the function in $\mathcal{L}^2_{\rho_X}$.

The following properties clearly hold :

- $\widetilde{K_x \otimes K_x}_{|\mathcal{H}} = K_x \otimes K_x$
- $\mathbb{E}\left(\widetilde{K_x \otimes K_x}\right) = \mathcal{T}$
- $\mathbb{E}\left(K_x \otimes K_x\right) = \Sigma$ as it has been noticed above.

For some $x \in \mathcal{X}$, we may denote $x \otimes x \;:= K_x \otimes K_x$. Moreover, abusing notations, we may forget the $\sim$ in many cases.

B.2. *Semi-stochastic recursion - intuition.* We remind that :

$$g_n = (I - \gamma K_{x_n} \otimes K_{x_n}) g_{n-1} + \gamma y_n K_{x_n},$$

with $\theta_0 = 0$. We have denoted $\Xi_n = (y_n - g_{\mathcal{H}}(x_n))K_{x_n}$. Thus $y_n K_{x_n} = g_{\mathcal{H}}(x_n)K_{x_n} + \Xi_n \stackrel{\text{def}}{=} \widetilde{K_{x_n} \otimes K_{x_n}} g_{\mathcal{H}} + \Xi_n$, and our recursion may be rewritten :

$$g_n - g_{\mathcal{H}} = (I - \gamma \widetilde{K_{x_n} \otimes K_{x_n}})(g_{n-1} - g_{\mathcal{H}}) + \gamma \Xi_n, \tag{B.3}$$

Finally, we are studying a sequence $(\eta_n)_n$ defined by :

$$\begin{aligned} \eta_0 &= g_{\mathcal{H}}, \\ \eta_n &= (I - \gamma_n \widetilde{K_{x_n} \otimes K_{x_n}})\eta_{n-1} + \gamma_n \Xi_n. \end{aligned} \tag{B.4}$$

Behaviour : It appears that to understand how this will behave, we may compare it to the following recursion, which may be described as a "semi-stochastic" version of (B.4) : we keep the randomness due to the noise Ξ_n but forget the randomness due to sampling by replacing $\widetilde{K_{x_n} \otimes K_{x_n}}$ by its expectation Σ (T, more precisely) :

$$\begin{aligned} \eta_0^{sto} &= g_{\mathcal{H}}, \\ \eta_n^{sto} &= (I - \gamma_n \Sigma)\eta_{n-1}^{sto} + \gamma_n \Xi_n. \end{aligned} \tag{B.5}$$

Complete proof : This comparison will give an interesting insight and the main terms of bias and variance will appear if we study (B.5). However this is not the true recursion : to get Theorem 2, we will have to do a bit of further work : we will first separate the error due to the noise from the error due to the initial condition, then link the true recursions to their "semi-stochastic" counterparts to make the variance and bias terms appear. That will be done in Section B.3.

Semi-stochastic recursion : In order to get such intuition, in both the finite horizon and on-line case, we will begin by studying the semi-stochastic equation (B.5).

First, we have, by induction:

$$
\begin{aligned}
\forall j \geqslant 1 \qquad \eta_j^{ssto} &= (I - \gamma_j \Sigma)\eta_{j-1}^{ssto} + \gamma_j \Xi_j. \\
\eta_j^{ssto} &= \left[\prod_{i=1}^{j}(I - \gamma_i \Sigma)\right] \eta_0^{ssto} + \sum_{k=1}^{j} \left[\prod_{i=k+1}^{j} (I - \gamma_i \Sigma)\right] \gamma_k \Xi_k \\
\eta_j^{ssto} &= D(1, j, (\gamma_i)_i)\eta_0^{ssto} + \sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k \\
\overline{\eta}_n^{ssto} &= \frac{1}{n}\sum_{j=1}^{n} D(1, j, (\gamma_i)_i)\eta_0^{ssto} + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{j} D(1, j, (\gamma_i)_i)\gamma_k \Xi_k.
\end{aligned}
$$

Then :

$$
\begin{aligned}
\mathbb{E}\|\overline{\eta}_n^{ssto}\|_{L^2_{\rho_X}}^2 &= \frac{1}{n^2}\mathbb{E}\|\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}} + \sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\|_{L^2_{\rho_X}} \\
&= \underbrace{\frac{1}{n^2}\,\mathbb{E}\|\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}}\|_{L^2_{\rho_X}}}_{\mathrm{Bias}(n)} \\
&+ \underbrace{2\frac{1}{n^2}\mathbb{E}\langle\sum_{j=1}^{n} D(1, j, (\gamma_i)_i) g_{\mathcal{H}}, \sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\rangle_{L^2_{\rho_X}}}_{=0 \text{ by (A.2)}\ ,} \\
&+ \underbrace{\frac{1}{n^2}\mathbb{E}\|\sum_{j=1}^{n}\sum_{k=1}^{j} D(k+1, j, (\gamma_i)_i)\gamma_k \Xi_k\|_{L^2_{\rho_X}}}_{\mathrm{Var}(n)} \qquad \text{(B.6)}
\end{aligned}
$$

In the following, all calculations may be driven either with $\|\Sigma^{1/2} \cdot\|_K$ or in $\|\cdot\|_{L^2_{\rho_X}}$ using the isometrical character of $\Sigma^{1/2}$. In order to simplify comparison with existing work and especially [13], we will mainly use the former as all calculations are only algebraic sums, we may sometimes use the notation $\langle x, \Sigma x\rangle_H$ instead of $\|\Sigma^{1/2}x\|_{\mathcal{H}}^2$. It is an abuse if $x \notin \mathcal{H}$, but however does not induce any confusion or mistake. In the following, if not explicitely specified, $\|\cdot\|$ will denote $\|\cdot\|_K$.

In the following we will thus denote :

$$\begin{aligned}\text{Bias}\left(n,(\gamma_i)_i,\Sigma,g_{\mathcal{H}}\right) &= \frac{1}{n^2}\mathbb{E}\left\|\Sigma^{1/2}\sum_{j=1}^{n}\left[\prod_{i=1}^{j}(I-\gamma_i\Sigma)\right]g_{\mathcal{H}}\right\|_K^2\\ \text{Var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right) &= \frac{1}{n^2}\mathbb{E}\left\|\Sigma^{1/2}\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\Xi_k\right\|_K^2.\end{aligned}$$

In section B.5 we will prove the following Lemmas which upper bound these bias and variance terms under different assumptions :

1. $\text{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)$ if we assume **A3,4**, γ constant,
2. $\text{Var}\left(n,\gamma,\Sigma,(\Xi_i)_i\right)$ if we assume **A3,6**, γ constant,
3. $\text{Bias}\left(n,(\gamma_i)_i,\Sigma,g_{\mathcal{H}}\right)$ if we assume **A3,4** and $\gamma_i=\frac{1}{n^\zeta}$, $0\leqslant\zeta\leqslant 1$,
4. $\text{Var}\left(n,(\gamma_i)_i,\Sigma,(\Xi_i)_i\right)$ if we assume **A3,6** and $\gamma_i=\frac{1}{n^\zeta}$, $0\leqslant\zeta\leqslant 1$.

The two terms show respectively the impact :

1. of the initial setting and the hardness to forget the initial condition,
2. the noise.

Thus the first one tends to decrease when γ is increasing, whereas the second one increases when γ increases. We understand we may have to choose our step γ in order to optimize the trade-off between these two factors.

In the finite-dimensional case, it results from such a decomposition that if $C=\sigma^2\Sigma$ then $\mathbb{E}\left[\langle\overline{\alpha}_{n-1},\Sigma\overline{\alpha}_{n-1}\rangle\right]\leqslant\frac{1}{n\gamma}\|\alpha\|_0^2+\frac{\sigma^2 d}{n}$, as this upper bound is vacuous when d is either large or infinite, we can derive comparable bounds in the infinite-dimensional setting under our assumptions **A3,4,6**.

Lemma 4 (Bias, **A3,4**, γ const.). *Assume **A3-4** and let α (resp. r) be the constant in **A3** (resp. **A4**) :*

***If** $r\leqslant 1$:*

$$\text{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)\leqslant\|\Sigma^{-r}g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2\left(\frac{1}{(n\gamma)^{2r}}\right)\overset{not}{=}\text{bias}(n,\gamma,r).$$

***If** $r\geqslant 1$:*

$$\text{Bias}\left(n,\gamma,\Sigma,g_{\mathcal{H}}\right)\leqslant\|\Sigma^{-r}g_{\mathcal{H}}\|_{L^2_{\rho_X}}^2\left(\frac{1}{n^2\gamma^r}\right)\overset{not}{=}\text{bias}(n,\gamma,r).$$

LEMMA 5 (Var, **A3,4**, γ const). *Assume **A3,6**, let α, s be the constants in **A3**, and σ the constant in **A6** (so that $\mathbb{E}\left[\Xi_n \otimes \Xi_n\right] \preccurlyeq \sigma^2\Sigma$).*

$$\mathrm{Var}\left(n, \gamma, \Sigma, (\Xi_i)_i\right) \leqslant C(\alpha)\ s^{2/\alpha}\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n} \overset{not}{=} \mathrm{var}(n, \gamma, \sigma^2, r, \alpha),$$

with $C(\alpha) = \frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$.

LEMMA 6 (Bias, **A3,4**, $(\gamma)_i$). *Assume **A3-4** and let α (resp. r) be the constant in **A3** (resp. **A4**). Assume we consider a sequence $\gamma_i = \frac{\gamma_0}{i^\zeta}$ with $0 < \zeta < 1$ then :*

1. *if $r(1-\zeta) < 1$:*

$$\begin{aligned} \mathrm{Bias}\left(n, (\gamma_i)_i, \Sigma, g_{\mathcal{H}}\right) &= O\left(\|\Sigma^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} n^{-2r(1-\zeta)}\right) \\ &= O\left(\|\Sigma^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \frac{1}{(n\gamma_n)^{2r}}\right), \end{aligned}$$

2. *if $r(1-\zeta) > 1$:*

$$\mathrm{Bias}\left(n, (\gamma_i)_i, \Sigma, g_{\mathcal{H}}\right) = O\left(\frac{1}{n^2}\right).$$

LEMMA 7 (Var, **A3,4**, $(\gamma)_i$). *Assume **A3,6**, let α, s be the constants in **A3**, and σ the constant in **A6** . If we consider a sequence $\gamma_i = \frac{\gamma_0}{i^\zeta}$ with $0 < \zeta < 1$ then :*

1. *if $0 < \zeta < \frac{1}{2}$ then*

$$\mathrm{Var}\left(n, (\gamma_i)_i, \Sigma, (\Xi_i)_i\right) = O\left(n^{-1+\frac{1-\zeta}{\alpha}}\right) = O\left(\frac{\sigma^2 (s^2\gamma_n)^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}\right),$$

2. *and if $\zeta > \frac{1}{2}$ then*

$$\mathrm{Var}\left(n, (\gamma_i)_i, \Sigma, (\Xi_i)_i\right) = O\left(n^{-1+\frac{1-\zeta}{\alpha}+2\zeta-1}\right).$$

Those Lemmas are proved in section B.5.

Considering decomposition (B.6) and our Lemmas above, we can state a first Proposition.

PROPOSITION 23 (Semi-stochastic recursion). *Assume **A1-6**. Let's consider the semi-stochastic recursion (that is the sequence :* $\eta_n = (I - \gamma_n\Sigma)\eta_{n-1} + \gamma_n\Xi_n$*) instead of our recursion initially defined. In the finite horizon setting, thus with* $\gamma_i = \gamma$ *for* $i \leqslant n$*, we have :*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_\rho)\right] \leqslant C(\alpha)\ s^{\frac{2}{\alpha}}\ \sigma^2 \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}} + \frac{\sigma^2}{n} + \|\Sigma^{-r} g_\rho\|^2_{L^2_{\rho_X}} \left(\frac{1}{n^{2\min\{r,1\}}\gamma^{2r}}\right).$$

Theorem 2 must be compared to Proposition 23 : Theorem 2 is just an extension but with the true stochastic recursion instead of the semi-stochastic one.

We finish this first part by a very simple Lemma which states that what we have done above is true for any semi stochastic recursion under few assumptions. Indeed, to get the complete bound, we will always come back to semi-stochastic type recursions, either without noise, or with a null initial condition.

LEMMA 8. *Let's assume:*

1. $\alpha_n = (I - \gamma\Sigma)\alpha_{n-1} + \gamma\Xi_n^\alpha$*, with* $\gamma\Sigma \preccurlyeq I$*.*
2. $(\Xi_n^\alpha) \in \mathcal{H}$ *is* $\mathcal{F}_n$ *measurable for a sequence of increasing* σ*-fields* $(\mathcal{F}_n)$*.*
3. $\mathbb{E}\left[\Xi_n^\alpha | \mathcal{F}_{n-1}\right] = 0$*,* $\mathbb{E}\left[\|\Xi_n^\alpha\|^2 | \mathcal{F}_{n-1}\right]$ *is finite and* $\mathbb{E}\left[\Xi_n^\alpha \otimes \Xi_n^\alpha\right] \preccurlyeq \sigma_\alpha^2\Sigma$*.*

Then :

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma\overline{\alpha}_{n-1}\rangle\right] = \text{Bias}\left(n, \gamma, \Sigma, \alpha_0\right) + \text{Var}\left(n, \gamma, \Sigma, (\Xi_i^\alpha)_i\right). \tag{B.7}$$

And we may apply Lemmas 4 and 5 if we have good assumptions on Σ, α_0*.*

B.3. *Complete proof, Theorem 2 (finite horizon) .* In the following, we will focus on the finite horizon setting, i.e., we assume the step size is constant, but may depend on the total number of observations n : for all $1 \leqslant i \leqslant n$, $\gamma_i = \gamma = \Gamma(n)$. The main idea of the proof is to be able to :

1. separate the different sources of error (noise & initial conditions),
2. then bound the difference between the stochastic recursions and their semi-stochastic versions, a case in which we are able to compute bias and variance as it is done above.

Our main tool will be the Minkowski's inequality, which is the triangular inequality for $\mathbb{E}\left(\|\cdot\|_{L^2_{\rho_X}}\right)$. This will allow us to separate the error due to the noise from the error due to the initial conditions. The sketch of the decomposition is given in Table 4.

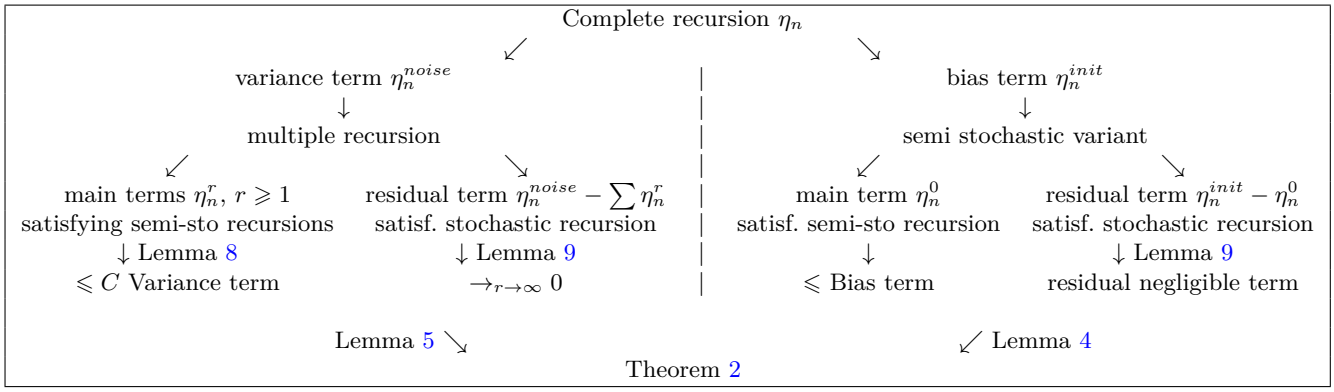


TABLE 4
Error decomposition in the finite horizon setting.

We remind that $(\eta_n)_n$ is defined by :

$$\eta_0 = g_{\mathcal{H}}, \text{ and the recusion } \eta_n = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1} + \gamma \Xi_n.$$

B.3.1. *A Lemma on stochastic recursions.* Before studying the main decomposition in Section B.3.2 we must give a classical Lemma on stochastic recursions which will be useful below :

LEMMA 9. *Assume $(x_n, \Xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\Xi_n | \mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[\|\Xi_n\|^2 | \mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\|K_{x_n}\|^2 K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] \preccurlyeq R^2\Sigma$, with $\mathbb{E}\left[K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] = \Sigma$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator Σ. Consider the recursion $\alpha_n = (I - \gamma K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma \Xi_n$, with $\gamma R^2 \leqslant 1$. Then :*

$$(1 - \gamma R^2)\, \mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] + \frac{1}{2n\gamma} \mathbb{E}\|\alpha_n\|^2 \leqslant \frac{1}{2n\gamma} \|\alpha_0\|^2 + \frac{\gamma}{n} \sum_{k=1}^{n} \mathbb{E}\|\Xi_k\|^2.$$

Especially, if $\alpha_0 = 0$, we have

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{(1 - \gamma R^2)} \frac{\gamma}{n} \sum_{k=1}^{n} \mathbb{E}\|\Xi_k\|^2.$$

Its proof may be found in [13] : it is a direct consequence of the classical recursion to upper bound $\|\alpha_n\|^2$.

B.3.2. *Main decomposition.*
We consider :

1. $(\eta_n^{init})_n$ defined by :
$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{init}.$$
η_n^{init} is the part of $(\eta_n)_n$ which is due to the **initial conditions** (it is equivalent to assuming $\Xi_n \equiv 0$).
2. Respectively, let $(\eta_n^{noise})_n$ be defined by :
$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise} + \gamma \Xi_n.$$
η_n^{noise} is the part of $(\eta_n)_n$ which is due to **the noise**.

A straightforward induction shows that for any n, $\eta_n = \eta_n^{init} + \eta_n^{noise}$ and $\bar{\eta}_n = \bar{\eta}_n^{init} + \bar{\eta}_n^{noise}$. Thus Minkowski's inequality, applied to $\left(\mathbb{E}\left[\|\cdot\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$, leads to :

$$\left(\mathbb{E}\left[\|\bar{\eta}_n\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\bar{\eta}_n^{init}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$$

(B.8)
$$\left(\mathbb{E}\left[\langle \bar{\eta}_n, \Sigma \bar{\eta}_n \rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma \bar{\eta}_n^{noise} \rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init} \rangle\right]\right)^{1/2}.$$

That means we can always consider separately the effect of the noise and the effect of the initial conditions. We'll first study η_n^{noise} and then η_n^{init}.

B.3.3. *Noise process.* We remind that $(\eta_n^{noise})_n$ is defined by :

$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise} + \gamma \Xi_n. \tag{B.9}$$

We are going to define some other sequences, which are defined by the following "semi-stochastic" recursion, in which $K_{x_n} \otimes K_{x_n}$ has been replaced be its expectancy Σ : first we define $\left(\eta_n^{noise,0}\right)_n$ so that

$$\eta_0^{noise,0} = 0 \text{ and } \eta_n^{noise,0} = (I - \gamma \Sigma)\eta_{n-1}^{noise,0} + \gamma \Xi_n.$$

Triangular inequality will allow us to upper bound $\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$:

(B.10)
$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,0}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} + \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise} - \bar{\eta}_n^{noise,0}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}$$

So that we're interested in the sequence $\left(\eta_n^{noise} - \eta_n^{noise,0}\right)_n$: we have

$$\begin{aligned}
\eta_0^{noise} - \eta_0^{noise,0} &= 0, \\
\eta_n^{noise} - \eta_n^{noise,0} &= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^0) + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^0 \\
&= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^0) + \gamma \Xi_n^1. \qquad \text{(B.11)}
\end{aligned}$$

which is the same type of Equation as (B.9). We have denoted $\Xi_n^1 = (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^0$.

Thus we may consider the following sequence, satisfying the "semi-stochastic" version of recursion (B.11), changing $K_{x_n} \otimes K_{x_n}$ into its expectation Σ : we define $\left(\eta_n^{noise,1}\right)_n$ so that:

$$\eta_0^{noise,1} = 0 \text{ and } \eta_n^{noise,1} = (I - \gamma\Sigma)\eta_{n-1}^{noise,1} + \gamma\Xi_n^1. \tag{B.12}$$

Thanks to the triangular inequality, we're interested in $\left(\eta_n^{noise} - \eta_n^{noise,0} - \eta_n^{noise,1}\right)_n$, which satisfies the (B.9)-type recursion :

$$\begin{aligned}
\eta_0^{noise} - \eta_0^{noise,0} - \eta_0^{noise,1} &= 0, \\
\eta_n^{noise} - \eta_n^{noise,0} - \eta_n^{noise,1} &= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{noise,0} - \eta_n^{noise,1}) \\
&\qquad + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,1} \\
&= (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_{n-1}^{noise} - \eta_{n-1}^{noise,0} - \eta_n^{noise,1}) + \gamma\Xi_n^{(2)}.
\end{aligned}$$

With $\Xi_n^{(2)} := (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,1}$.

And so on... For any $r \geqslant 0$ we define a sequence $(\eta_n^{noise,r})_n$ by :

$$\eta_0^{noise,r} = 0 \text{ and } \eta_n^{noise,r} = (I - \gamma\Sigma)\eta_{n-1}^{noise,r} + \gamma\Xi_n^r,$$
$$\text{with } \Xi_n^r = (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,r-1}.$$

We have, for any $r, n \in \mathbb{N}^2$:

$$\begin{aligned}
\eta_0^{noise} - \sum_{i=0}^{r} \eta_0^{noise,i} &= 0, \\
\eta_n^{noise} - \sum_{i=0}^{r} \eta_n^{noise,i} &= (I - \gamma K_{x_n} \otimes K_{x_n}) \left(\eta_{n-1}^{noise} - \sum_{i=0}^{r} \eta_{n-1}^{noise,i}\right) \\
&\qquad + \gamma(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{noise,r}. \\
&= (I - \gamma K_{x_n} \otimes K_{x_n}) \left(\eta_{n-1}^{noise} - \sum_{i=0}^{r} \eta_{n-1}^{noise,i}\right) + \gamma\Xi_n^{(r+1)}. \qquad \text{(B.13)}
\end{aligned}$$

So that $(\eta_n^{noise,r+1})$ follows the "semi-stochastic" version of (B.13)...

Minkowski's inequality. Considering this decomposition, we have, for any r, using triangular inequality :

(B.14)

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2} \leqslant \sum_{i=0}^{r} \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^{noise,i}\right\|_{L^2_{\rho_X}}^2\right]\right)^{1/2}$$

Moment Bounds. For any $i \geqslant 0$, we find that we may apply Lemma 8 to the sequence $(\eta_n^{noise,i})$. Indeed :

1. For any $r \geqslant 0$, $(\eta_n^{noise,r})$ is defined by :
$$\eta_0^{noise,r} = 0 \text{ and } \eta_n^{noise,r} = (I - \gamma\Sigma)\eta_{n-1}^{noise,r} + \gamma\Xi_n^r,$$
$$\text{with } \Xi_n^r = \begin{cases} (\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^{r-1} & \text{if } r \geqslant 1. \\ \Xi_n \qquad \text{if } r = 0. \end{cases}$$
2. for any $r \geqslant 0$, for all $n \geqslant 0$, Ξ_n^r is $\mathcal{F}_n := \sigma\left((x_i, z_i)_{1 \leqslant i \leqslant n}\right)$ measurable. (for $r = 0$ we use the definition of Ξ_n (**H4**), and by induction, for any $r \geqslant 0$ if we have $\forall n \in \mathbb{N}$, Ξ_n^r is $\mathcal{F}_n$ measurable, then for any $n \in \mathbb{N}$, by induction on n, $\eta_n^{noise,r}$ is $\mathcal{F}_n$ measurable, thus for any $n \in \mathbb{N}$, Ξ_n^{r+1} is $\mathcal{F}_n$ measurable.)
3. for any $r, n \geqslant 0$, $\mathbb{E}\left[\Xi_n | \mathcal{F}_{n-1}\right] = 0$: as shown above, η_{n-1}^{r-1} is $\mathcal{F}_{n-1}$ measurable so $\mathbb{E}\left[\ \Xi_n\ |\mathcal{F}_{n-1}\right] = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n} | \mathcal{F}_{n-1}\right] \eta_{n-1}^{noise,r-1} = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right] \eta_{n-1}^{noise,r-1} = 0$ (as x_n is independent of $\mathcal{F}_{n-1}$ by **A5** and $\mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right] = \mathbb{E}\left[\Sigma - K_{x_n} \otimes K_{x_n}\right]$ by **H4**).
4. $\mathbb{E}\left[\|\Xi_n^r\|^2\right]$ is finite (once again, by **A2** if $r = 0$ and by a double recursion to get the result for any $r, n \geqslant 0$).
5. We have to find a bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$. To do that, we are going, once again to use induction on r.

LEMMA 10. *For any $r \geqslant 0$ we have*
$$\begin{aligned} \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] &\preccurlyeq \gamma^r R^{2r} \sigma^2 \Sigma \\ \mathbb{E}\left[\eta_n^{noise,r} \otimes \eta_n^{noise,r}\right] &\preccurlyeq \gamma^{r+1} R^{2r} \sigma^2 I. \end{aligned}$$

Lemma 10. We make an induction on n.

Initialisation : for $r = 0$ we have by **A6** that $\mathbb{E}\left[\Xi_n^0 \otimes \Xi_n^0\right] \preccurlyeq \sigma^2\Sigma$. Moreover
$$\begin{aligned} \mathbb{E}(\eta_n^0 \otimes \eta_n^0) &= \gamma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{n-k} \mathbb{E}\left[\Xi_n^0 \otimes \Xi_n^0\right] (I - \gamma\Sigma)^{n-k} \\ &\preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{2(n-k)} \Sigma. \end{aligned}$$

We get
$$\forall n \geqslant 0, \quad \mathbb{E}\left[\eta_n^0 \otimes \eta_n^0\right] \preccurlyeq \gamma^2\sigma^2 \sum_{k=1}^{n-1} (I - \gamma\Sigma)^{2n-2-k} \Sigma \preccurlyeq \gamma\sigma^2 I.$$

Recursion : If we assume that for any $n \geqslant 0, \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] \preccurlyeq \gamma^r R^{2r}\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma^{r+1}R^{2r}\sigma^2 I$ then for any $n \geqslant 0$:

$$\begin{aligned}
\mathbb{E}\left[\Xi_n^{r+1} \otimes \Xi_n^{r+1}\right] &\preccurlyeq \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^r \otimes \eta_{n-1}^r(\Sigma - K_{x_n} \otimes K_{x_n})\right] \\
&= \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\mathbb{E}\left[\eta_{n-1}^r \otimes \eta_{n-1}^r\right](\Sigma - K_{x_n} \otimes K_{x_n})\right] \\
&\qquad\qquad (\text{as } \eta_{n-1} \in \mathcal{F}_{n-1}) \\
&\preccurlyeq \gamma^{r+1}R^{2r}\sigma^2\mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})^2\right] \\
&\preccurlyeq \gamma^{r+1}R^{2r+2}\sigma^2\Sigma.
\end{aligned}$$

Once again we have $(\eta_n^{r+1}) = \gamma^2 \sum_{k=1}^{n-1}(I - \gamma\Sigma)^{n-1-k}\Xi_n^{r+1}$, for any n:

$$\begin{aligned}
\mathbb{E}\left[\eta_n^{r+1} \otimes \eta_n^{r+1}\right] &\preccurlyeq \gamma^2\mathbb{E}\left[\sum_{k=1}^{n}(I - \gamma\Sigma)^{n-1-k}\Xi_n^{r+1} \otimes \Xi_n^{r+1}(I - \gamma\Sigma)^{n-1-k}\right] \\
&\preccurlyeq \gamma^{r+3}R^{2r+2}\sigma^2\sum_{k=1}^{n}(I - \gamma\Sigma)^{2n-2-2k}\Sigma \\
&\preccurlyeq \gamma^{r+2}R^{2r+2}\sigma^2 I.
\end{aligned}$$

□

With the bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$ and as we have said, with Lemma 8:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right] = \mathbb{E}\left[\langle\bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right] &\leqslant \operatorname{var}(n, \gamma, \sigma^2\gamma^i R^{2i}, s, \alpha) \\
&\leqslant \gamma^i R^{2i} \underset{\parallel}{\operatorname{var}}(n, \gamma, \sigma^2, s, \alpha) \,.. \qquad \text{(B.15)}
\end{aligned}$$

Moreover, using the Lemma on stochastic recursions (Lemma 9) for $(\bar{\eta}_n^{noise} - \sum_{i=0}^r \bar{\eta}_n^i)_n$ (all conditions are satisfied) we have :

$$\begin{aligned}
(1 - \gamma R^2)\ \mathbb{E}\left[\left\langle \bar{\eta}_n^{noise} - \sum_{i=0}^{r}\bar{\eta}_n^i, \Sigma\left(\bar{\eta}_n^{noise} - \sum_{i=0}^{r}\bar{\eta}_n^i\right)\right\rangle\right] &\leqslant \frac{\gamma}{n}\sum_{i=1}^{n}\mathbb{E}\|\Xi_k^{r+1}\|^2 \\
&\leqslant \gamma\operatorname{tr}\left(\mathbb{E}\left[\Xi_k^{r+1} \otimes \Xi_k^{r+1}\right]\right) \\
&\leqslant \gamma^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma) \\
\text{(B.16)} \quad \text{that is} \quad \mathbb{E}\left[\left\|\bar{\eta}_n^{noise} - \sum_{i=0}^{r}\bar{\eta}_n^{noise,i}\right\|_{L^2_{\rho_X}}^2\right] &\leqslant \gamma^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma).
\end{aligned}$$

Conclusion. Thus using (B.14), (B.15) and (B.16) :

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma \bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \left(\frac{1}{1-\gamma R^2}\gamma^{r+2}\sigma^2 R^{2r+2}\operatorname{tr}(\Sigma)\right)^{1/2} + \operatorname{var}(n,\gamma,\sigma^2,s,\alpha)^{1/2}\sum_{i=0}^{r}\left(\gamma R^2\right)^{i/2}. \tag{B.17}$$

And using the fact that $\gamma R < 1$, when $r \to \infty$ we get:

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{noise}, \Sigma \bar{\eta}_n^{noise}\rangle\right]\right)^{1/2} \leqslant \operatorname{var}(n,\gamma,\sigma^2,s,\alpha)^{1/2}\frac{1}{1-\sqrt{\gamma R^2}}. \tag{B.18}$$

Which is the main result of this part.

B.3.4. *Initial conditions.* We are now interested in getting such a bound for $\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init}\rangle\right]$. As this part stands for the initial conditions effect we may keep in mind that we would like to get an upper bound comparable to what we found for the Bias term in the proof of Proposition 1.

We remind that :

$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})\eta_{n-1}^{init}.$$

and define $(\eta_n^0)_{n\in\mathbb{N}}$ so that :

$$\eta_0^0 = g_{\mathcal{H}}, \quad \eta_n^0 = (I - \gamma\Sigma)\eta_{n-1}^0.$$

Minkowski's again. As above

$$\left(\mathbb{E}\left[\langle \bar{\eta}_n^{init}, \Sigma \bar{\eta}_n^{init}\rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle \bar{\eta}_n^{init} - \bar{\eta}_n^0, \Sigma\left(\bar{\eta}_n^{init} - \bar{\eta}_n^0\right)\rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0\rangle\right]\right)^{1/2}. \tag{B.19}$$

First for $\bar{\eta}_n^0$ we have a semi-stochastic recursion, with $\Xi_n \equiv 0$ so that we have

$$\mathbb{E}\langle \bar{\eta}_n^0, \Sigma \bar{\eta}_n^0\rangle \leqslant \operatorname{bias}(n,\gamma,r).$$

Then , for the residual term we use Lemma 9. Using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma K_{x_n} \otimes K_{x_n})(\eta_n^0 - \eta_n^{init}) + \gamma(K_{x_n} \otimes K_{x_n} - \Sigma)\eta_{n-1}^0,$$

we may apply **Lemma 9** to the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$ and $\Xi_n = (K_{x_n} \otimes K_{x_n} - \Sigma)\eta_{n-1}^0$. That is (as $\alpha_0 = 0$):

$$\mathbb{E}\langle \bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle \leqslant \frac{1}{1-\gamma R^2}\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^{n}\|\Xi_k\|^2\right]. \tag{B.20}$$

Now

$$\begin{aligned}\mathbb{E}\|\Xi_k\|^2 &= \mathbb{E}\left[\langle \eta_0, (I-\gamma\Sigma)^k(\Sigma - x_k\otimes x_k)^2(I-\gamma\Sigma)^k\eta_0\rangle\right]\\ &\leqslant \langle \eta_0, (I-\gamma\Sigma)^k R^2\Sigma(I-\gamma\Sigma)^k\eta_0\rangle\\ &\leqslant R^2\langle \eta_0, (I-\gamma\Sigma)^{2k}\Sigma\eta_0\rangle.\end{aligned}$$

Thus :

$$\begin{aligned}\frac{\gamma}{n}\mathbb{E}\left[\sum_{k=1}^n\|\Xi_k\|^2\right] &\leqslant \frac{\gamma R^2}{n}\langle\eta_0, \sum_{k=1}^n(I-\gamma\Sigma)^{2k}\Sigma\eta_0\rangle\\ &\leqslant \frac{\gamma R^2}{n}\left\|\left(\sum_{k=1}^n(I-\gamma\Sigma)^{2k}\Sigma^{2r}\right)^{1/2}\Sigma^{1/2-r}\eta_0\right\|^2\\ &\leqslant \frac{\gamma R^2}{n}\gamma^{-2r}\left|\!\left|\!\left|\sum_{k=1}^n(I-\gamma\Sigma)^{2k}(\gamma\Sigma)^{2r}\right|\!\right|\!\right| \|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}.\end{aligned}$$

$|||A^{1/2}|||^2 = |||A|||$. Moreover, as Σ is self adjoint, we have:

$$\begin{aligned}\left|\!\left|\!\left|\sum_{k=1}^n(I-\gamma\Sigma)^{2k}(\gamma\Sigma)^{2r}\right|\!\right|\!\right| &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^n(1-x)^{2k}(x)^{2r}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(1-x)^{2n}}{1-(1-x)^2}(x)^{2r}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(x)^{2n}}{1-x^2}(1-x)^{2r}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}\frac{1-(x)^{2n}}{1+x}(1-x)^{2r-1}\\ &\leqslant \sup_{0\leqslant x\leqslant 1}(1-(1-x)^{2n})(x)^{2r-1}\\ &\leqslant n^{1-2r}\end{aligned}$$

Where we have used inequality (B.44).
So that we would get, replacing our result in (B.20) :

$$\mathbb{E}\langle\bar\eta_n^0-\bar\eta_n^{noise}, \Sigma(\bar\eta_n^0-\bar\eta_n^{noise})\rangle \leqslant \frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}. \tag{B.21}$$

Conclusion. Summing both bounds we get from (B.19) :

$$\left(\mathbb{E}\left[\langle\bar\eta_n^{init}, \Sigma\bar\eta_n^{init}\rangle\right]\right)^{1/2} \leqslant \left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2} + (Bias(n,\gamma,g_{\mathcal{H}},\alpha))^{1/2}. \tag{B.22}$$

B.3.5. *Conclusion.* These two parts allow us to show Theorem 2 : using (B.22) and (B.18) in (B.8), and Lemmas 4 and 5 we have the final result.

Assuming **A1-6** :

1. If $r < 1$

$$\begin{aligned}\left(2\,\mathbb{E}\left[\varepsilon\left(g_n\right)-\varepsilon(g_{\mathcal{H}})\right]\right)^{1/2} \;\leqslant\; & \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\; s^{\frac{2}{\alpha}}\;\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}\right)^{1/2} \\ & +\left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\left(\frac{1}{(n\gamma)^{2r}}\right)\right)^{1/2} \\ & +\left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2}.\end{aligned}$$

2. If $r > 1$

$$\begin{aligned}\left(2\,\mathbb{E}\left[\varepsilon\left(g_n\right)-\varepsilon(g_{\mathcal{H}})\right]\right)^{1/2} \;\leqslant\; & \frac{1}{1-\sqrt{\gamma R^2}}\left(C(\alpha)\; s^{\frac{2}{\alpha}}\;\sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}\right)^{1/2} \\ & +\left(\|\Sigma^{-r}g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\left(\frac{1}{n^2\gamma^{2r}}\right)\right)^{1/2} \\ & +\left(\frac{1}{1-\gamma R^2}\frac{\gamma R^2}{(\gamma n)^{2r}}\|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}\right)^{1/2}.\end{aligned}$$

Regrouping terms, we get exactly Theorem 2. In order to derive corollaries, one just has to chose $\gamma = \Gamma(n)$ in order to balance the main terms.

B.4. *Complete proof, Theorem 3 (on-line setting).* The sketch of the proof is exactly the same. We just have to check that changing a constant step into a decreasing sequence of step-size does not change to much. However as most calculations make appear some weird constants, we will only look for asymptotics. The sketch of the decomposition is given in Table 5.

B.4.1. *A Lemma on stochastic recursions - on-line.* We want to derive a Lemma comparable to Lemma 9 in the online setting. That is considering a sequence $(\gamma_n)_n$ and the recursion $\alpha_n = (I - \gamma_n K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma_n \Xi_n$ we would like to have a bound on $\mathbb{E}\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1}\rangle$.

LEMMA 11. *Assume $(x_n, \Xi_n) \in \mathcal{H}\times\mathcal{H}$ are $\mathcal{F}_n$ measurable for a sequence of increasing σ-fields $(\mathcal{F}_n)$. Assume that $\mathbb{E}\left[\Xi_n|\mathcal{F}_{n-1}\right] = 0$, $\mathbb{E}\left[\|\Xi_n\|^2|\mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\|K_{x_n}\|^2 K_{x_n}\otimes K_{x_n}|\mathcal{F}_{n-1}\right] \preccurlyeq R^2\Sigma$, with $\mathbb{E}\left[K_{x_n}\otimes K_{x_n}|\mathcal{F}_{n-1}\right] =$*

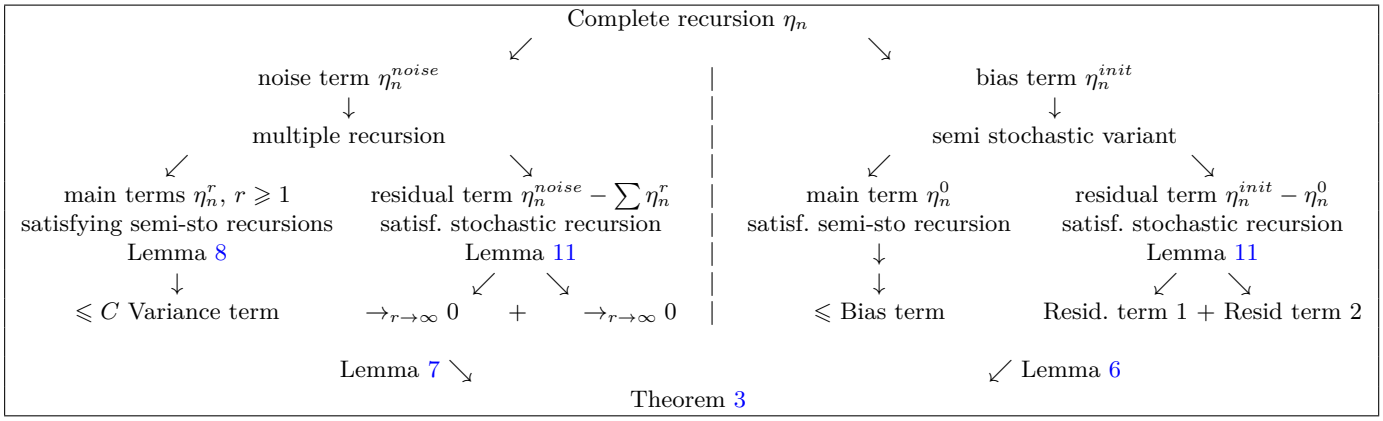


TABLE 5
Sketch of the proof, on-line setting.

Σ *for all* $n \geqslant 1$ *, for some* $R > 0$ *and invertible operator* Σ*. Consider the recursion* $\alpha_n = (I - \gamma_n K_{x_n} \otimes K_{x_n})\alpha_{n-1} + \gamma_n \Xi_n$*, with* $(\gamma_n)_n$ *a sequence such that for any* n*,* $\gamma_n R^2 \leqslant 1$*. Then if* $\alpha_0 = 0$*, we have So that if* $\alpha_0 = 0$ *:*

(B.23)
$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\sum_{i=1}^{n-1} \|\alpha_i\|^2 \left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right) + \sum_{k=1}^{n} \gamma_k \mathbb{E}\|\Xi_k\|^2 \right).$$

PROOF.

(B.24) $$2\gamma_n(1-\gamma_n R^2)\mathbb{E}\langle \Sigma \alpha_{n-1}, \alpha_{n-1} \rangle \leqslant \mathbb{E}\left(\|\alpha_{n-1}\|^2 - \|\alpha_n\|^2 + \gamma_n^2 \|\Xi_n\|^2 \right)$$

So that, if we assume that (γ_n) is non increasing:

(B.25) $$\mathbb{E}\langle \Sigma \alpha_{n-1}, \alpha_{n-1} \rangle \leqslant \frac{1}{2\gamma_n(1-\gamma_0 R^2)} \mathbb{E}\left(\|\alpha_{n-1}\|^2 - \|\alpha_n\|^2 + \gamma_n^2 \|\Xi_n\|^2 \right)$$

Using convexity :

$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\frac{\|\alpha_0\|^2}{\gamma_1} + \sum_{i=1}^{n-1} \|\alpha_i\|^2 \underbrace{\left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right)}_{\geqslant 0} - \frac{\|\alpha_n\|^2}{\gamma_n} + \sum_{k=1}^{n} \gamma_k \mathbb{E}\|\Xi_k\|^2 \right).$$

So that if $\alpha_0 = 0$:

(B.26)
$$\mathbb{E}\left[\langle \overline{\alpha}_{n-1}, \Sigma \overline{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2n(1-\gamma_0 R^2)} \left(\sum_{i=1}^{n-1} \|\alpha_i\|^2 \left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}} \right) + \sum_{k=1}^{n} \gamma_k \mathbb{E}\|\Xi_k\|^2 \right).$$

Note that it may be interesting to consider the weighted average $\tilde{\alpha}_n = \frac{\sum \gamma_i \alpha_i}{\sum \gamma_i}$, which would satisfy be convexity

(B.27)
$$\mathbb{E}\left[\langle \tilde{\alpha}_{n-1}, \Sigma \tilde{\alpha}_{n-1} \rangle\right] \leqslant \frac{1}{2(\sum \gamma_i)(1-\gamma_0 R^2)} \left(\frac{\|\alpha_0\|^2}{\gamma_1} - \frac{\|\alpha_n\|^2}{\gamma_n} + \sum_{k=1}^{n} \gamma_k^2 \mathbb{E}\|\Xi_k\|^2 \right).$$

□

B.4.2. *Noise process.* We remind that $(\eta_n^{noise})_n$ is defined by :

$$\eta_0^{noise} = 0 \text{ and } \eta_n^{noise} = (I - \gamma K_{x_n} \otimes K_{x_n}) \eta_{n-1}^{noise} + \gamma \Xi_n. \tag{B.28}$$

As before, for any $r \geqslant 0$ we define a sequence $(\eta_n^{noise,r})_n$ by :

$$\eta_0^{noise,r} = 0 \text{ and } \eta_n^{noise,r} = (I - \gamma \Sigma) \eta_{n-1}^{noise,r} + \gamma \Xi_n^r,$$

$$\text{with } \Xi_n^r = (\Sigma - K_{x_n} \otimes K_{x_n}) \eta_{n-1}^{noise,r-1}.$$

And we want to use the following upper bound

(B.29)
$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \sum_{i=0}^{r} \left(\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\bar{\eta}_n^{noise} - \sum_{i=0}^{r} \bar{\eta}_n^{noise,i}\right\|^2_{L^2_{\rho_X}}\right]\right)^{1/2}.$$

So that we had to upper bound the noise :

Lemma 12. *For any $r \geqslant 0$ we have $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] \preccurlyeq R^{2r} \gamma_0^r \sigma^2 \Sigma$ and $\mathbb{E}\left[\eta_n^{noise,r} \otimes \eta_n^{noise,r}\right] \preccurlyeq \gamma_0^{r+1} R^{2r} \sigma^2 I$.*

Lemma 12. We make an induction on n.
We note that :

$$\begin{aligned}
\sum_{k=1}^{n} D(n, k+1, (\gamma_k)_k) \gamma_k^2 \Sigma D(n, k+1, (\gamma_k)_k) &\leqslant \gamma_0 \sum_{k=1}^{n} D(n, k+1, (\gamma_k)_k) \gamma_k \Sigma \\
&\leqslant \gamma_0 \sum_{k=1}^{n} D(n, k+1, (\gamma_k)_k) - D(n, k, (\gamma_k)_k) \\
&\leqslant \gamma_0 (I - D(n, 1, (\gamma_k)_k)) \\
&\leqslant \gamma_0 I \qquad \text{(B.30)}
\end{aligned}$$

Where we have used that : $D(n, k+1, (\gamma_k)_k) - D(n, k, (\gamma_k)_k) = D(n, k+1, (\gamma_k)_k) \gamma_k \Sigma$.

Initialisation : for $r=0$ we have by **A6** that $\mathbb{E}\left[\Xi_n^0 \otimes \Xi_n^0\right] \preccurlyeq \sigma^2\Sigma$. Moreover $\eta_n^0 = \sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k\Xi_k^0$.

$$\begin{aligned}
\mathbb{E}(\eta_n^0 \otimes \eta_n^0) &= \sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k^2\mathbb{E}\left[\Xi_k^0 \otimes \Xi_k^0\right] D(k+1,n,(\gamma_k)_k) \\
&\preccurlyeq \sigma^2 \sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k^2\Sigma D(k+1,n,(\gamma_k)_k) \\
&\preccurlyeq \sigma^2\gamma_0 I, \quad \text{by (B.30)}
\end{aligned}$$

Induction : If we assume $\forall n \geqslant 0, \quad \mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right] \preccurlyeq \gamma_0^r R^{2r}\sigma^2\Sigma$ and $\mathbb{E}\left[\eta_n^r \otimes \eta_n^r\right] \preccurlyeq \gamma_0^{r+1}R^{2r}\sigma^2 I$ then: $\forall n \geqslant 0$,

$$\begin{aligned}
\mathbb{E}\left[\Xi_n^{r+1} \otimes \Xi_n^{r+1}\right] &\preccurlyeq \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\eta_{n-1}^r \otimes \eta_{n-1}^r(\Sigma - K_{x_n} \otimes K_{x_n})\right] \\
&= \mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})\mathbb{E}\left[\eta_{n-1}^r \otimes \eta_{n-1}^r\right](\Sigma - K_{x_n} \otimes K_{x_n})\right] \\
&\qquad\qquad (\text{as } \eta_{n-1} \in \mathcal{F}_{n-1}) \\
&\preccurlyeq \gamma_0^{r+1}R^{2r}\sigma^2\mathbb{E}\left[(\Sigma - K_{x_n} \otimes K_{x_n})^2\right] \\
&\preccurlyeq \gamma_0^{r+1}R^{2r+2}\sigma^2\Sigma.
\end{aligned}$$

Once again we have $\eta_n^{r+1} = \sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k\Xi_k^{r+1}$, for any n:

$$\begin{aligned}
\mathbb{E}\left[\eta_n^{r+1} \otimes \eta_n^{r+1}\right] &\preccurlyeq \gamma^2\mathbb{E}\left[\sum_{k=1}^n (I-\gamma\Sigma)^{n-1-k}\Xi_n^{r+1} \otimes \Xi_n^{r+1}(I-\gamma\Sigma)^{n-1-k}\right] \\
&\preccurlyeq \sigma^2\gamma_0^{r+1}R^{2r}\sum_{k=1}^n D(n,k+1,(\gamma_k)_k)\gamma_k^2\Sigma D(k+1,n,(\gamma_k)_k) \\
&\preccurlyeq \sigma^2\gamma_0^{r+2}R^{2r}I, \quad \text{by (B.30)}
\end{aligned}$$

□

With the bound on $\mathbb{E}\left[\Xi_n^r \otimes \Xi_n^r\right]$ and as we have said, with Lemma 8:
(B.31)
$\mathbb{E}\left[\|\bar{\eta}_n^{noise,i}\|_{L^2_{\rho_X}}^2\right] = \mathbb{E}\left[\langle\bar{\eta}_n^i, \Sigma\bar{\eta}_n^i\rangle\right] \leqslant \mathrm{var}(n,\gamma,\alpha,\gamma_0^i R^{2i}\sigma,s) = \gamma_0^i R^{2i}\,\mathrm{var}(n,\gamma,\alpha,\sigma,s)$.

Moreover, using the Lemma on stochastic recursions (Lemma 11) for $(\alpha_n^r)_n = (\eta_n^{noise} - \sum_{i=0}^r \eta_n^i)_n$ (all conditions are satisfied) we have :

$$2(1-\gamma_0 R^2)\,\mathbb{E}\left[\left\langle\overline{\alpha}_n^r, \Sigma\overline{\alpha}_n^r\right\rangle\right] \leqslant \frac{1}{n}\left(\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i^r\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right)+\sum_{k=1}^n\gamma_k\mathbb{E}\|\Xi_k^{r+1}\|^2\right).$$

We are going to show that both these terms goes to 0 when r goes to infinity. Indeed :

$$\begin{aligned}\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k^{r+1}\|^2 &\leqslant \sum_{k=1}^{n}\gamma_k \operatorname{tr}\left(\mathbb{E}\left[\Xi_k^{r+1}\otimes\Xi_k^{r+1}\right]\right)\\ &\leqslant \sum_{k=1}^{n}\gamma_k\gamma_0^{r+1}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma)\\ &\leqslant n\gamma_0^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma)\end{aligned}$$

Moreover, if we assume $\gamma_i = \frac{1}{i^\zeta}$:

$$\frac{1}{n}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i^r\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right)\leqslant 2\zeta\frac{1}{n}\sum_{i=1}^{n-1}\frac{\gamma_i}{i}\mathbb{E}\|\alpha_i^r\|^2$$

And

$$\alpha_i^r = (I-\gamma_i\widetilde{K_{x_i}\otimes K_{x_i}})\alpha_{i-1}^r+\gamma_i\Xi_i$$

So that :

$$\begin{aligned}\|\alpha_i^r\| &\leqslant \|\|(I-\gamma_i\widetilde{K_{x_i}\otimes K_{x_i}})\|\|\,\|\alpha_{i-1}^r\|+\gamma_i\,\|\Xi_i\|\\ &\leqslant \|\alpha_{i-1}^r\|+\gamma_i\,\|\Xi_i\|\\ &\leqslant \sum_{k=1}^{i}\gamma_k\,\|\Xi_k\|.\\ \text{thus : } \|\alpha_i^r\|^2 &\leqslant \sum_{k=1}^{i}\gamma_k\sum_{k=1}^{i}\gamma_k\,\|\Xi_k\|^2\\ \mathbb{E}\|\alpha_i^r\|^2 &\leqslant \sum_{k=1}^{i}\gamma_k\sum_{k=1}^{i}\gamma_k\,\mathbb{E}\|\Xi_k\|^2\\ \mathbb{E}\|\alpha_i^r\|^2 &\leqslant C_1\;\; i\gamma_i\;\; i\gamma_0^{r+2}R^{2r+2}\sigma^2\operatorname{tr}(\Sigma)\\ \frac{\gamma_i}{i}\,\mathbb{E}\|\alpha_i^r\|^2 &\leqslant C_2\;\; i\gamma_i^2\;\; (\gamma_0R^2)^{r+2}\\ \frac{1}{n}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i^r\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right) &\leqslant C_3\;\; n\gamma_n^2\;\; (\gamma_0R^2)^{r+2}.\end{aligned}$$

That is :

$$E\left[\left\|\bar{\eta}_n^{noise}-\sum_{i=0}^{r}\bar{\eta}_n^{noise,i}\right\|^2_{L^2_{\rho_X}}\right]\leqslant(\gamma_0R^2)^{r+2}\left(\sigma^2\operatorname{tr}(\Sigma)+C_3n\gamma_n^2\right). \tag{B.32}$$

With (B.29), (B.31),(B.32), we get :

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \sum_{i=0}^{r}\left(\gamma_0^i R^{2i}\,\mathrm{var}(n,\gamma,\alpha,\sigma,s)\right)^{1/2} + \left((\gamma_0 R^2)^{r+2}\left(\sigma^2\,\mathrm{tr}(\Sigma) + C_3\; n\gamma_n^2\right)\right)^{1/2}. \tag{B.33}$$

So that, with $r \to \infty$:

$$\left(\mathbb{E}\left[\|\bar{\eta}_n^{noise}\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant \left(C\,\mathrm{var}(n,\gamma,\alpha,\sigma,s)\right)^{1/2}. \tag{B.34}$$

B.4.3. *Initial conditions.* Exactly as before, we can separate the effect of initial conditions and of noise : We are interested in getting such a bound for $\mathbb{E}\left[\langle\bar{\eta}_n^{init}, \Sigma\bar{\eta}_n^{init}\rangle\right]$. We remind that :

$$\eta_0^{init} = g_{\mathcal{H}} \text{ and } \eta_n^{init} = (I - \gamma_n K_{x_n}\otimes K_{x_n})\eta_{n-1}^{init}.$$

and define $(\eta_n^0)_{n\in\mathbb{N}}$ so that :

$$\eta_0^0 = g_{\mathcal{H}}, \quad \eta_n^0 = (I - \gamma_n\Sigma)\eta_{n-1}^0.$$

Minkowski's again :. As above

$$\left(\mathbb{E}\left[\langle\bar{\eta}_n^{init}, \Sigma\bar{\eta}_n^{init}\rangle\right]\right)^{1/2} \leqslant \left(\mathbb{E}\left[\langle\bar{\eta}_n^{init} - \bar{\eta}_n^0, \Sigma\left(\bar{\eta}_n^{init} - \bar{\eta}_n^0\right)\rangle\right]\right)^{1/2} + \left(\mathbb{E}\left[\langle\bar{\eta}_n^0, \Sigma\bar{\eta}_n^0\rangle\right]\right)^{1/2}. \tag{B.35}$$

First for $\bar{\eta}_n^0$ we have a semi-stochastic recursion, with $\Xi_n \equiv 0$ so that we have

$$\langle\bar{\eta}_n^0, \Sigma\bar{\eta}_n^0\rangle \leqslant \mathrm{Bias}(n,(\gamma_n)_n, g_{\mathcal{H}}, r). \tag{B.36}$$

Then , for the residual term we use **Lemma** 11 for the recursion above with $\alpha_n = \eta_n^0 - \eta_n^{init}$. Using that :

$$\eta_n^0 - \eta_n^{init} = (I - \gamma K_{x_n}\otimes K_{x_n})(\eta_n^0 - \eta_n^{init}) + \gamma_n(K_{x_n}\otimes K_{x_n} - \Sigma)\eta_{n-1}^0,$$

That is (as $\alpha_0 = 0$):

$$\mathbb{E}\langle\bar{\eta}_n^0 - \bar{\eta}_n^{noise}, \Sigma(\bar{\eta}_n^0 - \bar{\eta}_n^{noise})\rangle \leqslant \frac{1}{2n(1-\gamma_0 R^2)}\left(\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i} + \frac{1}{\gamma_{i+1}}\right) + \sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k\|^2\right). \tag{B.37}$$

Now

$$\begin{aligned}\mathbb{E}\|\Xi_k\|^2 &= \mathbb{E}\left[\langle \eta_0, D(n,1,(\gamma_i)_i)(\Sigma - x_k\otimes x_k)^2 D(1,n,(\gamma_i)_i)\eta_0\rangle\right]\\ &\leqslant R^2\langle \eta_0, D(1,n,(\gamma_i)_i)^2\Sigma\eta_0\rangle.\end{aligned}$$

Thus :

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^n \gamma_k\|\Xi_k\|^2\right] &\leqslant R^2\langle \eta_0, \sum_{k=1}^n \gamma_k D(1,n,(\gamma_i)_i)^2\Sigma\eta_0\rangle\\ &\leqslant R^2\left\|\left(\sum_{k=1}^n \gamma_k D(1,n,(\gamma_i)_i)^2\Sigma^{2r}\right)^{1/2}\Sigma^{1/2-r}\eta_0\right\|^2\\ &\leqslant R^2\left|\left|\left|\sum_{k=1}^n D(1,n,(\gamma_i)_i)^2\gamma_k\Sigma^{2r}\right|\right|\right| \|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}. \qquad \text{(B.38)}\end{aligned}$$

Now :

$$\begin{aligned}\left|\left|\left|\sum_{k=1}^n D(1,n,(\gamma_i)_i)^2\gamma_k\Sigma^{2r}\right|\right|\right| &\leqslant \sup_{0\leqslant x\leqslant 1/\gamma_0}\sum_{k=1}^n\prod_{i=1}^n(1-\gamma_i x)^2\gamma_k x^{2r}\\ &\leqslant \sup_{0\leqslant x\leqslant 1/\gamma_0}\sum_{k=1}^n \exp\left(-\sum_{i=1}^k\gamma_i x\right)\gamma_k x^{2r}\\ &\leqslant \sup_{0\leqslant x\leqslant 1/\gamma_0}\sum_{k=1}^n \exp\left(-k\gamma_k x\right)\gamma_k x^{2r} \quad \text{if } (\gamma_k)_k \text{ is decreasing}\\ &\leqslant \gamma_0\sup_{0\leqslant x\leqslant 1/\gamma_0}\sum_{k=1}^n \exp\left(-k\gamma_k x\right)x^{2r}\\ &\leqslant \gamma_0\sup_{0\leqslant x\leqslant 1/\gamma_0}\sum_{k=1}^n \exp\left(-k^{1-\rho}\gamma_0 x\right)x^{2r} \quad \text{if } (\gamma_k)_i = \frac{\gamma_0}{k^\rho}\\ &\leqslant \gamma_0\sup_{0\leqslant x\leqslant 1/\gamma_0} x^{2r}\int_{u=0}^n \exp\left(-u^{1-\rho}\gamma_0 x\right)du\end{aligned}$$

$$\int_{u=0}^{n-1}\exp\left(-u^{1-\rho}\gamma_0 x\right)du \leqslant n \quad \text{clearly, but also}$$

$$\int_{u=0}^{n-1}\exp\left(-u^{1-\rho}\gamma_0 x\right)du \leqslant \int_{t=0}^{\infty}\exp\left(-t^{1-\rho}\right)(x\gamma_0)^{-\frac{1}{1-\rho}}dt \quad \text{changing variables. So that :}$$

$$\begin{aligned}\left|\left|\left|\sum_{k=1}^n D(1,n,(\gamma_i)_i)^2\gamma_k\Sigma^{2r}\right|\right|\right| &\leqslant \gamma_0\sup_{0\leqslant x\leqslant 1/\gamma_0} x^{2r}\left(n\wedge I(x\gamma_0)^{-\frac{1}{1-\rho}}\right)\\ &\leqslant \gamma_0 C_1\sup_{0\leqslant x\leqslant 1/\gamma_0}\left(nx^{2r}\wedge x^{2r-\frac{1}{1-\rho}}\right) \text{ and if } 2r-\frac{1}{1-\rho}<0\\ &\leqslant \gamma_0 C_1 n^{1-2r(1-\rho)}. \qquad \text{(B.39)}\end{aligned}$$

And finally, using (B.38), (B.39) :

$$\begin{aligned}\frac{1}{2n(1-\gamma_0 R^2)}\sum_{k=1}^{n}\gamma_k\mathbb{E}\|\Xi_k\|^2 &\leqslant \frac{\gamma_0 C_1\ \|\Sigma^{-r}\eta_0\|^2_{L^2_{\rho_X}}R^2}{2(1-\gamma_0 R^2)}(n\gamma_n)^{-2r}\\ &\leqslant K(n\gamma_n)^{-2r}. \qquad \text{(B.40)}\end{aligned}$$

To conclude, we have to upper bound :

$$\frac{1}{2n(1-\gamma_0 R^2)}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right).$$

By the induction we make to get **Lemma** 11, we have :

$$\begin{aligned}\|\alpha_i\|^2 &\leqslant \|\alpha_{i-1}\|^2+\gamma_i^2\|\Xi_i\|^2\\ &\leqslant \sum_{k=1}^{i}\gamma_k^2\|\Xi_k\|^2\\ &\leqslant \sum_{k=1}^{i}\gamma_k\|\Xi_k\|^2\\ &\leqslant Ci\ (i\gamma_i)^{-2r}.\end{aligned}$$

So that (C changes during calculation) :

$$\begin{aligned}\frac{1}{2n(1-\gamma_0 R^2)}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\left(-\frac{1}{\gamma_i}+\frac{1}{\gamma_{i+1}}\right) &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1}\mathbb{E}\|\alpha_i\|^2\frac{\gamma_i}{i}\\ &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1}i\ (i\gamma_i)^{-2r}\frac{\gamma_i}{i}\\ &\leqslant C\frac{1}{n}\sum_{i=1}^{n-1}(i\gamma_i)^{-2r}\gamma_i\\ &\leqslant C\frac{\gamma_n}{(n\gamma_n)^{2r}}.\end{aligned}$$

So that we would get, replacing our result in (B.37) :

$$\mathbb{E}\langle\bar{\eta}_n^0-\bar{\eta}_n^{noise},\Sigma(\bar{\eta}_n^0-\bar{\eta}_n^{noise})\rangle = O\left(\frac{1}{n\gamma_n}\right)^{2r}+O\left(\frac{\gamma_n}{n\gamma_n}\right)^{2r}=O\left(\frac{1}{n\gamma_n}\right)^{2r}. \qquad \text{(B.41)}$$

And finally, with (B.36) and (B.41) in (B.35),

$$\begin{aligned}\left(\mathbb{E}\left[\langle\bar{\eta}_n^{init},\Sigma\bar{\eta}_n^{init}\rangle\right]\right)^{1/2} &\leqslant \left(\mathbb{E}\left[\langle\bar{\eta}_n^{init}-\bar{\eta}_n^0,\Sigma\left(\bar{\eta}_n^{init}-\bar{\eta}_n^0\right)\rangle\right]\right)^{1/2}+\left(\mathbb{E}\left[\langle\bar{\eta}_n^0,\Sigma\bar{\eta}_n^0\rangle\right]\right)^{1/2}\\ &\leqslant \left(O\left(\frac{1}{n\gamma_n}\right)^{2r}\right)^{1/2}+\text{bias}(n,(\gamma_n)_n,g_{\mathcal{H}},r)^{1/2}. \qquad \text{(B.42)}\end{aligned}$$

B.4.4. *Conclusion.* We conclude with both (B.34) and (B.42) in (B.8) :

(B.43)
$$\left(\mathbb{E}\left[\|\bar{\eta}_n\|^2_{L^2_{\rho_X}}\right]\right)^{1/2} \leqslant (C\,\mathrm{var}(n,\gamma,\alpha,\sigma,s))^{1/2} + \left(O\left(\frac{1}{n\gamma_n}\right)^{2r}\right)^{1/2} + \mathrm{bias}(n,(\gamma_n)_n,g_{\mathcal{H}},r)^{1/2}.$$

Which gives Theorem 3 using Lemmas 6 and 7. Once again, deriving corollaries is simple.

B.5. *Some quantities.* In this section, we bound the main quantities which are involved above.

B.5.1. *Lemma 4.*

Lemma 4.

If $0 \leqslant r \leqslant 1$:

$$\begin{aligned}
\mathrm{Bias}(n,\gamma,g_{\mathcal{H}},r) &= \frac{1}{n^2}\langle \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k g_{\mathcal{H}}, \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \ \ \Sigma g_{\mathcal{H}}\rangle \\
&= \frac{1}{n^2}\langle \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \Sigma^{2r}\Sigma^{-r+1/2} g_{\mathcal{H}}, \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \ \ \Sigma^{-r+1/2} g_{\mathcal{H}}\rangle \\
&= \frac{1}{n^2}\left\| \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \Sigma^r (\Sigma^{-r+1/2} g_{\mathcal{H}})\right\|^2 \\
&\leqslant \frac{1}{n^2}\left\|\left\| \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \Sigma^r \right\|\right\|^2 \ \ \left\|\Sigma^{-r+1/2} g_{\mathcal{H}}\right\|^2 \\
&= \frac{1}{n^2}\gamma^{-2r}\left\|\left\| \sum_{k=0}^{n-1}(I-\gamma\Sigma)^k \gamma^r\Sigma^r \right\|\right\|^2 \ \ \left\|\Sigma^{-r} g_{\mathcal{H}}\right\|^2_{\mathcal{L}^2_\rho} \\
&\leqslant \frac{1}{n^2}\gamma^{-2r} \sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right)^2 \left\|\Sigma^{-r} g_{\mathcal{H}}\right\|^2_{\mathcal{L}^2_\rho} \\
&\leqslant \left(\frac{1}{(n\gamma)^{2r}}\right)\left\|\Sigma^{-r} g_{\mathcal{H}}\right\|^2_{L^2_{\rho_X}}.
\end{aligned}$$

Using the inequality :

(B.44)
$$\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) \leqslant n^{1-r}.$$

Indeed :

$$\begin{aligned}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) &= \frac{1-(1-x)^n}{x}x^r \\ &= (1-(1-x)^n)x^{r-1}.\end{aligned}$$

And we have, for any $n \in \mathbb{N}, r \in [\,0;1], x \in [\,0;1] : (1-(1-x)^n) \leqslant (nx)^{1-r}$:

1. if $nx \leqslant 1$ then $(1-(1-x)^n) \leqslant nx \leqslant (nx)^{1-r}$ (the first inequality can be proved by deriving the difference).
2. if $nx \geqslant 1$ then $(1-(1-x)^n) \leqslant 1 \leqslant (nx)^{1-r}$.

If $r \geqslant 1$, $x \mapsto (1-(1-x)^n)$ is increasing on $[\,0;1]$ so $\sup_{0\leqslant x\leqslant 1}\left(\sum_{k=0}^{n-1}(1-x)^k x^r\right) = 1$: there is no improvement in comparison to $r = 1$:

$$\text{Bias}(n,\gamma,g_{\mathcal{H}},r) \leqslant \left(\frac{1}{n^2\gamma^{2r}}\right)\left|\left|\Sigma^{-r}g_{\mathcal{H}}\right|\right|^2_{L^2_{\rho_X}}.$$

□

B.5.2. *Lemma 5.*

Lemma 5 .
In the following proof, we consider $s = 1$. It's easy to get the complete result replacing in the proof below " γ" by " $s^2\gamma$". We have, for $j \in \mathbb{N}$, still assuming $\gamma\Sigma \preccurlyeq I$, and by a comparison to the integral :

$$\begin{aligned}\text{tr}\left(I-(I-\gamma\Sigma)^j\right)^2\Sigma^{-1}C &= \sigma^2\,\text{tr}\left(I-(I-\gamma\Sigma)^j\right)^2 \\ &\leqslant 1+\sigma^2\int_{u=1}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \\ &\quad \text{(1 stands for the first term in the sum)} \\ &= 1+\sigma^2\int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du \\ &\quad +\sigma^2\int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du.\end{aligned}$$

Note that the first integral may be empty if $\gamma j \leqslant 1$. We also have:

$$\text{tr}\left(I-(I-\gamma\Sigma)^j\right)^2\Sigma^{-1}C \geqslant \sigma^2\int_{u=1}^{\infty}\left(1-\left(1-\frac{\gamma}{u^\alpha}\right)^j\right)^2 du.$$

Considering that $g_j : u \mapsto \left(1 - (1 - \frac{\gamma}{u^\alpha})^j\right)^2$ is a decreasing function of u we get :

$$\forall u \in [1; (\gamma j)^{\frac{1}{\alpha}}], \quad (1 - e^{-1})^2 \leqslant g_j(u) \leqslant 1.$$

Where we have used the fact that $\left(1 - \frac{1}{j}\right)^j \leqslant e^{-1}$ for the left hand side inequality. Thus we have proved :

$$(1 - e^{-1})^2 (\gamma j)^{\frac{1}{\alpha}} \leqslant \int_{u=1}^{(\gamma j)^{\frac{1}{\alpha}}} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du \leqslant (\gamma j)^{\frac{1}{\alpha}}.$$

For the other part of the sum, we consider $h_j : u \mapsto \left(\frac{1-\left(1-\frac{\gamma}{u^\alpha}\right)^j}{\frac{\gamma}{u^\alpha}}\right)^2$ which is an increasing function of u. So :

$$\forall u \in [(\gamma j)^{\frac{1}{\alpha}}; +\infty], \quad (1 - e^{-1})^2 j^2 \leqslant h_j(u) \leqslant j^2,$$

using the same trick as above. Thus :

$$\begin{aligned}
\int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du &= \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} h_j(u) \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{\gamma}{u^\alpha}\right)^2 du \\
&\leqslant j^2 \gamma^2 \int_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \left(\frac{1}{u^\alpha}\right)^2 du \\
&= j^2 \gamma^2 \left[\frac{1}{(1-2\alpha) u^{2\alpha-1}}\right]_{u=(\gamma j)^{\frac{1}{\alpha}}}^{\infty} \\
&= j^2 \gamma^2 \frac{1}{(2\alpha-1)((\gamma j)^{\frac{1}{\alpha}})^{2\alpha-1}} \\
&= \frac{1}{(2\alpha-1)} (j\gamma)^{\frac{1}{\alpha}}.
\end{aligned}$$

And we could get, by a similar calculation :

$$\int_{u=(\gamma j)^{\frac{1}{\alpha}}+1}^{\infty} \left(1 - \left(1 - \frac{\gamma}{u^\alpha}\right)^j\right)^2 du \geqslant (1 - e^{-1})^2 \frac{1}{(2\alpha-1)} (j\gamma)^{\frac{1}{\alpha}}.$$

Finally, we have shown that :

$$C_1 (j\gamma)^{\frac{1}{\alpha}} \leqslant \operatorname{tr}\left(I - (I - \gamma\Sigma)^j\right)^2 \leqslant C_2 (j\gamma)^{\frac{1}{\alpha}} + 1.$$

Where $C_1 = (1-e^{-1})^2\ (1+\frac{1}{(2\alpha-1)})$ and $C_2 = (1+\frac{1}{(2\alpha-1)})$ are real constants.
To get the complete variance term we have to calculate : $\frac{\sigma^2}{n^2}\sum_{j=1}^{n-1}\operatorname{tr}\left(I-(I-\gamma\Sigma)^j\right.$.
We have :

$$\begin{aligned}
\frac{\sigma^2}{n^2}\sum_{j=1}^{n-1}\operatorname{tr}\left(I-(I-\gamma\Sigma)^j\right)^2 &\leqslant \frac{\sigma^2}{n^2}\sum_{j=1}^{n-1}\left(C_2(j\gamma)^{\frac{1}{\alpha}}+1\right)\\
&\leqslant \frac{\sigma^2}{n^2}C_2\gamma^{\frac{1}{\alpha}}\int_{u=2}^{n}u^{\frac{1}{\alpha}}du+\frac{\sigma^2}{n}\\
&\leqslant \frac{\sigma^2}{n^2}C_2\gamma^{\frac{1}{\alpha}}\frac{\alpha}{\alpha+1}n^{\frac{\alpha+1}{\alpha}}+\frac{\sigma^2}{n}\\
&\leqslant \frac{\alpha\ \sigma^2\ C_2}{\alpha+1}\ \frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n}.
\end{aligned}$$

That is :

$$(1-e^{-1})^2\ C(\alpha)\ \sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{(n-1)^{1-\frac{1}{\alpha}}} \leqslant \operatorname{Var}(n,\gamma,\alpha,\sigma^2) \leqslant C(\alpha)\ \sigma^2\frac{\gamma^{\frac{1}{\alpha}}}{n^{1-\frac{1}{\alpha}}}+\frac{\sigma^2}{n},$$

with $C(\alpha)=\frac{2\alpha^2}{(\alpha+1)(2\alpha-1)}$.

□

B.5.3. *Lemma 6.*

Proof.

$$\begin{aligned}
\frac{1}{n^2}\left\|\Sigma^{1/2}\sum_{k=1}^{n}\prod_{i=1}^{k}(I-\gamma_i\Sigma)\,g_{\mathcal{H}}\right\|_K^2 &\leqslant \frac{1}{n^2}\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}(I-\gamma_i\Sigma)\,\Sigma^r\right|\right|\right|^2||\Sigma^{1/2-r}g_{\mathcal{H}}||_K^2\\
&\leqslant \frac{1}{n^2}\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}(I-\gamma_i\Sigma)\,\Sigma^r\right|\right|\right|^2||\Sigma^{-r}g_{\mathcal{H}}||_{L^2_{\rho_X}}^2.
\end{aligned}$$

Moreover :

$$\begin{aligned}
\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right| &\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i x\right)x^r \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-\sum_{i=1}^{k}\gamma_i x\right)\gamma_k x^r \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-k\gamma_k x\right)\gamma_k x^r \quad \text{if } (\gamma_k)_k \text{ is decreasing} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-k\gamma_k x\right)x^r \\
&\leqslant \sup_{0\leqslant x\leqslant 1}\sum_{k=1}^{n}\exp\left(-k^{1-\zeta}x\right)x^r \quad \text{if } (\gamma_k)_i=\frac{1}{k^\zeta} \\
&\leqslant \sup_{0\leqslant x\leqslant 1}x^r\int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du \quad \text{by comparison to the integral} \\
\int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du &\leqslant n \quad \text{clearly, but also} \\
\int_{u=0}^{n}\exp\left(-u^{1-\zeta}x\right)du &\leqslant \int_{t=0}^{\infty}\exp\left(-t^{1-\zeta}\right)(x)^{-\frac{1}{1-\zeta}}dt \quad \text{changing variables. So that :} \\
\left|\left|\left|\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma^r\right|\right|\right| &\leqslant K\sup_{0\leqslant x\leqslant 1}x^r\left(n\wedge x^{-\frac{1}{1-\zeta}}\right) \\
&\leqslant K\sup_{0\leqslant x\leqslant 1}\left(nx^r\wedge x^{r-\frac{1}{1-\zeta}}\right) \text{ and if } r-\frac{1}{1-\zeta}<0 \\
&\leqslant Kn^{1-r(1-\zeta)}.
\end{aligned}$$

So that :

$$\begin{aligned}
\frac{1}{n^2}\left\langle\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)g_{\mathcal{H}},\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma g_{\mathcal{H}}\right\rangle &\leqslant \frac{1}{n^2}\left(Kn^{1-r(1-\zeta)}\right)^2||\Sigma^{-r}g_{\mathcal{H}}||^2_{L^2_{\rho_X}} \\
&\leqslant K^2||\Sigma^{-r}g_{\mathcal{H}}||^2_{L^2_{\rho_X}}n^{-2r(1-\zeta)}.
\end{aligned}$$

Else if $r-\frac{1}{1-\zeta}>0$, then $\sup_{0\leqslant x\leqslant 1}\left(nx^r\wedge x^{r-\frac{1}{1-\zeta}}\right)=1$, so that

$$\frac{1}{n^2}\left\langle\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)g_{\mathcal{H}},\sum_{k=1}^{n}\prod_{i=1}^{k}\left(I-\gamma_i\Sigma\right)\Sigma g_{\mathcal{H}}\right\rangle = O\left(\frac{||\Sigma^{-r}g_{\mathcal{H}}||^2_{L^2_{\rho_X}}}{n^2}\right).$$

□

B.5.4. *Lemma 7.*

PROOF. To get corollary 7, we will just replace in the following calculations γ by $s^2\gamma$ We remind that :

(B.45)
$$\mathrm{Var}\left(n,(\gamma_i)_i,\Sigma,(\xi_i)_i\right)=\frac{1}{n^2}\mathbb{E}\left\langle\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k,\Sigma\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k\right\rangle.$$

For shorter notation, in the following proof, we note $\mathrm{Var}(n)=\mathrm{Var}\left(n,(\gamma_i)_i,\Sigma,(\xi_i)_i\right)$.

$$\begin{aligned}
\mathrm{Var}(n) &= \frac{1}{n^2}\mathbb{E}\left\langle\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k,\Sigma\sum_{j=1}^{n}\sum_{k=1}^{j}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\mathbb{E}\left\langle\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k,\Sigma\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left\langle\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k,\Sigma\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\gamma_k\xi_k\right\rangle\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\left\langle M_{n,k}\gamma_k\xi_k,\Sigma M_{n,k}\gamma_k\xi_k\right\rangle \qquad \text{with } M_{n,k}:=\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\ \mathbb{E}\left\langle M_{n,k}\xi_k,\Sigma M_{n,k}\xi_k\right\rangle=\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\ \mathbb{E}\,\mathrm{tr}\left(M_{n,k}\Sigma M_{n,k}\xi_k\otimes\xi_k\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\,\mathrm{tr}\left(M_{n,k}^2\Sigma\Sigma\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\,\mathrm{tr}\left(\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}(I-\gamma_i\Sigma)\right]\right)\Sigma\right)^2\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\left(\left(\sum_{j=k}^{n}\left[\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right)\right]\right)\frac{1}{t^\alpha}\right)^2.
\end{aligned}$$

Let's first upper bound:

$$\begin{aligned}
\left[\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right)\right] &\leqslant \exp\sum_{i=k+1}^{j}(\gamma_i\frac{1}{t^\alpha}) \\
&= \exp-\sum_{i=k+1}^{j}\left(\frac{1}{i^\zeta}\frac{1}{t^\alpha}\right) \text{ if } \gamma_i=\frac{1}{i^\zeta} \\
&\leqslant \exp-\frac{1}{t^\alpha}\int_{u=k+1}^{j+1}\left(\frac{1}{u^\zeta}du\right) \\
&\leqslant \exp-\frac{1}{t^\alpha}\frac{(j+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}.
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{j=k}^{n}\prod_{i=k+1}^{j}\left(1-\gamma_i\frac{1}{t^\alpha}\right) &\leqslant \sum_{j=k}^{n}\exp-\frac{1}{t^\alpha}\frac{(j+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta} \\
&\leqslant \int_{u=k}^{n}\exp-\frac{1}{t^\alpha}\frac{(u+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du \\
&\leqslant (n-k) \quad \text{clearly}
\end{aligned}$$

(this upper bound is good when $t >> n^{1-\zeta}$), but we also have:

$$\int_{u=k}^{n}\exp-\frac{1}{t^\alpha}\frac{(u+1)^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du = \int_{u=k+1}^{n+1}\exp-\frac{1}{t^\alpha}\frac{u^{1-\zeta}-(k+1)^{1-\zeta}}{1-\zeta}du.$$

With $\rho = 1-\zeta, K_\zeta := \frac{1}{(1-\zeta)^{1/\rho} t^{\alpha/\rho}}$ and

$$\begin{aligned}
v^\rho &= \frac{1}{t^\alpha} \frac{(u)^\rho - (k+1)^\rho}{(1-\zeta)} \\
v &= \frac{1}{(1-\zeta)^{1/\rho} t^{\alpha/\rho}} \left((u)^\rho - (k+1)^\rho\right)^{1/\rho} \\
dv &= K_\zeta \frac{1}{\rho} \left(u^\rho - (k+1)^\rho\right)^{1/\rho - 1} \rho u^{\rho-1} du \\
dv &= K_\zeta \left(1 - \left(\frac{k+1}{u}\right)^\rho\right)^{1/\rho-1} du \\
dv \frac{1}{K_\zeta \left(1 - \left(\frac{(k+1)^\rho}{t^\alpha C v^\rho + (k+1)^\rho}\right)\right)^{1/\rho-1}} &= du \\
dv \frac{1}{K_\zeta} \left(\frac{t^\alpha C v^\rho + (k+1)^\rho}{t^\alpha C v^\rho + (k+1)^\rho - (k+1)^\rho}\right)^{1/\rho-1} &= du \\
dv \frac{1}{K_\zeta} \left(\frac{t^\alpha C v^\rho + (k+1)^\rho}{t^\alpha C v^\rho}\right)^{1/\rho-1} &= du \\
dv \frac{1}{K_\zeta} \left(1 + \frac{(k+1)^\rho}{t^\alpha C v^\rho}\right)^{1/\rho-1} &= du
\end{aligned}$$

$$\begin{aligned}
\int_{u=k}^{n} \exp - \frac{1}{t^\alpha} \frac{(u+1)^{\frac{\alpha}{\alpha+\beta}} - (k+1)^{\frac{\alpha}{\alpha+\beta}}}{(1-\zeta)} du &\leqslant \int_0^\infty \frac{1}{K_\zeta} \left(1 + \frac{(k+1)^\rho}{t^\alpha C v^\rho}\right)^{1/\rho-1} \exp\left(-v^\rho\right) dv \\
&\leqslant \frac{2^{1/\rho-1}}{K_\zeta} \int_0^\infty \left(1 \vee \frac{(k+1)^\rho}{t^\alpha C v^\rho}\right)^{1/\rho-1} \exp\left(-v^\rho\right) dv \\
&\leqslant 2^{1/\rho-1} (1-\zeta)^{1/\rho} t^{\alpha/\rho} \int_0^\infty \left(1 \vee \frac{(k+1)^{1-\rho}}{(t^\alpha C)^{1/\rho-1} v^{1-\rho}}\right) \exp\left(-v^\rho\right) dv. \\
&\leqslant K t^{\alpha/\rho} \left(I_1 \vee I_2 \frac{(k+1)^{1-\rho}}{(t^\alpha)^{1/\rho-1}}\right) \\
&\leqslant K \left(t^{\frac{\alpha}{1-\zeta}} \vee t^\alpha (k+1)^\zeta\right).
\end{aligned}$$

Finally :

$$
\begin{aligned}
\mathrm{Var}(n) &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)\wedge K\left(t^{\frac{\alpha}{1-\zeta}}\vee t^{\alpha}(k+1)^{\zeta}\right)\right)^2\\
\mathrm{Var}(n) &\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge K\left(t^{2\frac{\alpha}{1-\zeta}}+t^{2\alpha}k^{2\zeta}\right)\right)\\
&\leqslant \underbrace{\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge K\left(t^{2\frac{\alpha}{1-\zeta}}\right)\right)}_{S_1}\\
&\qquad\qquad + \underbrace{\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge t^{2\alpha}k^{2\zeta}\right)}_{S_2}\\
S_1 &\leqslant K\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{(n-k)^{\frac{1-\zeta}{\alpha}}}\frac{1}{t^{2\alpha}}t^{2\frac{\alpha}{1-\zeta}}+\sum_{t=(n-k)^{\frac{1-\zeta}{\alpha}}}^{\infty}\frac{1}{t^{2\alpha}}(n-k)^2\right)\\
&\leqslant K\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{(n-k)^{\frac{1-\zeta}{\alpha}}}t^{\frac{2\alpha\zeta}{1-\zeta}}+(n-k)^2\sum_{t=(n-k)^{\frac{1-\zeta}{\alpha}}}^{\infty}\frac{1}{t^{2\alpha}}\right)\\
&\leqslant G\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left((n-k)^{\frac{1-\zeta}{\alpha}(\frac{2\alpha\zeta}{1-\zeta}+1)}+(n-k)^2\frac{1}{(n-k)^{\frac{1-\zeta}{\alpha}(2\alpha-1)}}\right)\\
&\leqslant G\frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left((n-k)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}+(n-k)^{2-\frac{1-\zeta}{\alpha}(2\alpha-1)}\right)\\
&= 2G\sigma^2\frac{1}{n^2}\sum_{k=1}^{n}\frac{1}{k^{2\zeta}}(n-k)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\\
&\leqslant 2G\sigma^2\frac{1}{n^2}\sum_{k=1}^{n}\left(\frac{n}{k}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}k^{\frac{1-\zeta}{\alpha}}\\
&= 2G\sigma^2n^{-1+\frac{1-\zeta}{\alpha}}\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\left(\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\\
&= 2G\sigma^2n^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{\frac{(2\alpha-1)\zeta+1}{\alpha}}\left(\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\right)\\
&= 2G\sigma^2n^{-1+\frac{1-\zeta}{\alpha}}\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{k/n}-1\right)^{2\zeta}\left(1-\frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}}\right).
\end{aligned}
$$

If $\zeta < \frac{1}{2}$ then

$$\int_0^1 \left(\frac{1}{x} - 1\right)^{2\zeta} (1-x)^{\frac{1-\zeta}{\alpha}} dx < \infty$$

and

$$\begin{aligned} S_1 &\leqslant Hn^{-1+\frac{1-\zeta}{\alpha}} \left(\frac{1}{n} \sum_{k=1}^{n} \left(\frac{1}{k/n} - 1\right)^{2\zeta} \left(1 - \frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}} \right) \\ &\leqslant H' n^{-1+\frac{1-\zeta}{\alpha}}. \end{aligned}$$

If $\zeta > \frac{1}{2}$ then

$$\int_0^1 \left(\frac{1}{x} - 1\right)^{2\zeta} (1-x)^{\frac{1-\zeta}{\alpha}} - \left(\frac{1}{x}\right)^{2\zeta} dx < \infty.$$

and

$$\begin{aligned} S_1 &\leqslant Hn^{-1+\frac{1-\zeta}{\alpha}} \left(\frac{1}{n} \sum_{k=1}^{n} \left(\frac{1}{k/n} - 1\right)^{2\zeta} \left(1 - \frac{k}{n}\right)^{\frac{1-\zeta}{\alpha}} - \left(\frac{n}{k}\right)^{2\zeta} + \frac{1}{n} \sum_{k=1}^{n} \left(\frac{n}{k}\right)^{2\zeta} \right) \\ &\leqslant Hn^{-1+\frac{1-\zeta}{\alpha}} \left(C + n^{2\zeta-1} \right) \\ &\leqslant Cn^{-1+\frac{1-\zeta+\alpha(2\zeta-1)}{\alpha}}. \end{aligned}$$

$$
\begin{aligned}
S_2 &= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\sum_{t=1}^{\infty}\frac{1}{t^{2\alpha}}\left((n-k)^2\wedge t^{2\alpha}k^{2\zeta}\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(\sum_{t=1}^{t_\ell}\frac{1}{t^{2\alpha}}t^{2\alpha}k^{2\zeta}+\sum_{t=t_\ell}^{\infty}\frac{1}{t^{2\alpha}}(n-k)^2\right)\quad\text{with}\quad t_\ell=\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}}\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta}\sum_{t=1}^{t_\ell}1+(n-k)^2\sum_{t=t_\ell}^{\infty}\frac{1}{t^{2\alpha}}\right)\\
&\leqslant \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta}\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}}+(n-k)^2\left(\frac{(n-k)^{\frac{1}{\alpha}}}{k^{\frac{\zeta}{\alpha}}}\right)^{1-2\alpha}\right)\\
&= \frac{1}{n^2}\sum_{k=1}^{n}\gamma_k^2\sigma^2\left(k^{2\zeta-\frac{\zeta}{\alpha}}(n-k)^{\frac{1}{\alpha}}+(n-k)^{\frac{1}{\alpha}}k^{\frac{\zeta}{\alpha}(2\alpha-1)}\right)\\
&= \frac{2\sigma^2}{n^2}\sum_{k=1}^{n}\frac{1}{k^{2\zeta}}(n-k)^{\frac{1}{\alpha}}k^{\frac{\zeta}{\alpha}(2\alpha-1)}\\
&= \frac{2\sigma^2}{n^2}\sum_{k=1}^{n}k^{-\frac{\zeta}{\alpha}}(n-k)^{\frac{1}{\alpha}}\\
&= 2\sigma^2 n^{\left(-1+-\frac{\zeta}{\alpha}+\frac{1}{\alpha}\right)}\frac{1}{n}\sum_{k=1}^{n}\left(\frac{k}{n}\right)^{-\frac{\zeta}{\alpha}}\left(1-\frac{k}{n}\right)^{\frac{1}{\alpha}}\\
&\leqslant Kn^{\left(-1+\frac{1-\zeta}{\alpha}\right)}.
\end{aligned}
$$

As we have a Riemann sum which converges.

Finally we get : if $0<\zeta<\frac{1}{2}$ then

$$
\begin{aligned}
\mathrm{Var}(n) &= O\left(\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}}\right)\\
&= O\left(\sigma^2\frac{\sigma^2(s^2\gamma_n)1/\alpha}{n^{1-1/\alpha}}n^{-1+\frac{1-\zeta}{\alpha}}\right)
\end{aligned}
$$

where we have re-used the constants s by formaly replacing in the proof γ by γs^2.

and if $\zeta>\frac{1}{2}$ then

$$
\mathrm{Var}(n)=O\left(\sigma^2 n^{-1+\frac{1-\zeta}{\alpha}+2\zeta-1}\right).
$$

Which is substantially Lemma 7.

□

References.

[1] G. Wahba, Spline Models for observationnal data. SIAM, 1990.

[2] B. Schölkopf and A. J. Smola, Learning with Kernels. MIT Press, 2002.

[3] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

[4] S. Smale and F. Cucker, "On the mathematical foundations of learning," Bulletin of the American Mathematical Society, vol. 39, no. 1, pp. 1–49, 2001.

[5] A. Caponnetto and E. De Vito, "Optimal Rates for the Regularized Least-Squares Algorithm," Foundations of Computational Mathematics, vol. 7, no. 3, pp. 331–368, 2007.

[6] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," Proceedings of the International Conference on Learning Theory (COLT), 2012.

[7] L. Rosasco, A. Tacchetti, and S. Villa, "Regularization by Early Stopping for Online Learning Algorithms," ArXiv e-prints, 2014.

[8] Y. Ying and M. Pontil, "Online gradient descent learning algorithms," Foundations of Computational Mathematics, vol. 5, 2008.

[9] P. Tarrès and Y. Yao, "Online learning as stochastic approximation of regularization paths," ArXiv e-prints 1103.5538, 2011.

[10] H. Robbins and S. Monro, "A stochastic approxiation method," The Annals of mathematical Statistics, vol. 22, no. 3, pp. 400–407, 1951.

[11] S. Shalev-Shwartz, "Online learning and online convex optimization," Foundations and Trends in Machine Learning, vol. 4, no. 2, pp. 107–194, 2011.

[12] F. Bach and E. Moulines, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in Adv. NIPS, 2011.

[13] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n)," Advances in Neural Information Processing Systems (NIPS), 2013.

[14] A. Berlinet and C. Thomas-Agnan, Reproducing kernel Hilbert spaces in probability and statistics, vol. 3. Springer, 2004.

[15] B. Thomson, J. Bruckner, and A. M. Bruckner, Elementary real analysis. Pearson, 2000.

[16] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," Constructive Approximation, vol. 26, no. 2, pp. 153–172, 2007.

[17] H. Brezis, Analyse fonctionnelle, Théorie et applications. Masson, 1983.

[18] A. N. Kolmogorov and S. V. Fomin, Elements of the theory of functions and functional analysis, vol. 1. Courier Dover Publications, 1999.

[19] P. Mikusinski and E. Weiss, "The Bochner Integral," ArXiv e-prints, 2014.

[20] N. Aronszajn, "Theory of reproducing kernels," Transactions of the American Mathematical Society, vol. 68, no. 3, pp. 337–404, 1950.

[21] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," The Journal of Machine Learning Research, vol. 7, pp. 2651–2667, 2006.

[22] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet, "Universality, characteristic kernels and rkhs embedding of measures," The Journal of Machine Learning Research, vol. 12, pp. 2389–2410, 2011.

[23] J. Vert, "Kernel Methods ," 2014.

[24] D. Hsu, S. M. Kakade, and T. Zhang, "Random design analysis of ridge regression," Foundations of Computational Mathematics, vol. 14, no. 3, pp. 569–600, 2014.

[25] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," Constructive Approximation, vol. 26, no. 2, pp. 289–315, 2007.

[26] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," Journal of Mathematical Analysis and Applications, vol. 33, no. 1, pp. 82–95, 1971.
[27] G. H. Golub and C. F. V. Loan, Matrix Computations. J. Hopkins Univ. Press, 1996.
[28] M. W. Mahoney, "Randomized algorithms for matrices and data," Foundations and Trends in Machine Learning, vol. 3, no. 2, pp. 123–224, 2011.
[29] C. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in Adv. NIPS, 2001.
[30] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in Proceedings of the International Conference on Learning Theory (COLT), 2013.
[31] O. Dekel, S. Shalev-Shwartz, and Y. Singer, "The Forgetron: A kernel-based perceptron on a fixed budget," in Adv. NIPS, 2005.
[32] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," Journal of Machine Learning Research, vol. 6, pp. 1579–1619, 2005.
[33] J. Kivinen, S. A.J., and R. C. Williamson, "Online learning with kernels," IEEE transactions on signal processing, vol. 52, no. 8, pp. 2165–2176, 2004.
[34] Y. Yao, A dynamic Theory of Learning. PhD thesis, University of California at Berkeley, 2006.
[35] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," ICML 2014 Proceedings of the twenty-first international conference on machine learning, 2004.
[36] E. Hazan and S. Kale, "Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization," Proceedings of the International Conference on Learning Theory (COLT), 2011.
[37] H. W. Engl, M. Hanke, and N. A., "Regularization of inverse problems," Klüwer Academic Publishers, 1996.
[38] S. Lacoste-Julien, M. Schmidt, and F. Bach, "A simpler approach to obtaining an O(1/t) rate for the stochastic projected subgradient method," ArXiv e-prints 1212.2002, 2012.
[39] I. M. Johnstone, "Minimax Bayes, asymptotic minimax and sparse wavelet priors," Statistical Decision Theory and Related Topics, pp. 303–326, 1994.
[40] M. Abramowitz and I. Stegun, Handbook of mathematical functions. Dover publications, 1964.
[41] F. Paulin, Topologie, analyse et calcul différentiel. Notes de cours, École Normale Supérieure, 2009.
[42] H. Hochstadt, Integral equations. 1973.

SIERRA Project-Team
23, avenue d'Italie
75013 Paris, France
E-mail: aymeric.dieuleveut@ens.fr
francis.bach@ens.fr